Philipp Niemann
Vanessa van den Bogaert
Ricarda Ziegler *Hrsg.*

Evaluationsmethoden der Wissenschaftskommunikation





Evaluationsmethoden der Wissenschaftskommunikation

Philipp Niemann ·
Vanessa van den Bogaert · Ricarda Ziegler
(Hrsg.)

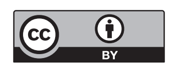
Evaluationsmethoden der Wissenschaftskommunikation



Hrsg.
Philipp Niemann
NaWik gGmbH
Heidelberg, Deutschland

Vanessa van den Bogaert
Lehrstuhl für Lehr- Lernforschung
Ruhr-Universität Bochum
Bochum, Deutschland

Ricarda Ziegler
Bereich Qualität & Transfer
Wissenschaft im Dialog gGmbH
Berlin, Deutschland



ISBN 978-3-658-39581-0 ISBN 978-3-658-39582-7 (eBook)
https://doi.org/10.1007/978-3-658-39582-7

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Barbara Emig-Roller
Springer VS ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Danksagung

Der herzliche Dank der Herausgeber:innen gilt dem **Bundesministerium für Bildung und Forschung** und der **Klaus Tschira Stiftung.**

Durch die großzügige Übernahme der Open-Access-Veröffentlichungskosten wird es möglich, dass dieses Buch, das sich explizit an Akteur:innen aus Praxis und Forschung der Wissenschaftskommunikation richtet, nicht nur einem wissenschaftlichen Publikum an Forschungs- und Bildungseinrichtungen zugänglich ist, sondern von vielen Menschen genutzt werden kann.

Inhaltsverzeichnis

Herausgeber- und Autorenverzeichnis

Über die Herausgeber

Philipp Niemann ist stellvertretender Direktor und wissenschaftlicher Leiter des Nationalen Instituts für Wissenschaftskommunikation (NaWik). Zuvor war er als Nachwuchsgruppenleiter im Department für Wissenschaftskommunikation am Karlsruher Institut für Technologie (KIT) tätig. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftskommunikation, qualitative Rezeptionsforschung und politische Kommunikation.

Vanessa van den Bogaert ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Lehr-Lernforschung der Ruhr-Universität Bochum. Sie widmet sich in ihren Forschungsschwerpunkten der wissenschaftlichen Begleitforschung von Citizen-Science-Projekten sowie der Grundlagenforschung zur Interessengenese an außerschulischen Lernorten. Sie leitet die Arbeitsgruppe *Science of Citizen Science* in Zusammenarbeit mit *Bürger schaffen Wissen*.

Ricarda Ziegler ist Leiterin des Bereichs Qualität & Transfer bei Wissenschaft im Dialog (WiD) – der deutschen Organisation für Wissenschaftskommunikation. Sie verantwortet dort u. a. die Impact Unit, die sich Fragen der Wirkung und Evaluation von Wissenschaftskommunikation widmet. Außerdem leitet sie das bevölkerungsrepräsentative Wissenschaftssurvey Wissenschaftsbarometer. Ricarda Ziegler hat einen Hintergrund in der Politikwissenschaft.

Autorenverzeichnis

Ferdinand Abacioglu M.Sc. IU Internationale Hochschule, Frankfurt am Main, Deutschland

Jacob Birkenhäger ifok GmbH, Berlin, Deutschland

Till Bruckermann Institut für Erziehungswissenschaft, Leibniz Universität Hannover, Hannover, Deutschland

Christoph Böhmert IU Internationale Hochschule, Karlsruhe, Deutschland

Bonnie Dietermann Museum für Naturkunde – Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Deutschland

Liliann Fischer Wissenschaft im Dialog, Berlin, Deutschland

Jens Fleischer Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland

Markus Gabriel com.X Institut, Bochum, Deutschland,

Rüdiger Goldschmidt Dialogik gemeinnützige Gesellschaft für Kommunikations- und Kooperationsforschung mbH, Stuttgart, Deutschland

Hannah Greving Arbeitsgruppe Wissenskonstruktion, Leibniz-Institut für Wissensmedien, Tübingen, Deutschland

Imke Hedder Wissenschaft im Dialog, Berlin, Deutschland

Armin Hempel SFB 980 „Episteme in Bewegung", Freie Universität Berlin, Berlin, Deutschland

Christian Humm M.A. Büro des Universitätspräsidenten, Universität des Saarlandes, Saarbrücken, Deutschland

Isabella Kessel Translake, Konstanz, Deutschland,

Sabrina H. Kessler Institut für Kommunikationswissenschaft und Medienforschung der Universität Zürich, Universität Zürich, Zürich, Schweiz

Joachim Kimmerle Arbeitsgruppe Wissenskonstruktion, Leibniz-Institut für Wissensmedien, Tübingen, Deutschland

Valerie Knapp Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland

Moritz Krell Didaktik der Biologie, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Deutschland

Julia Metag Institut für Kommunikationswissenschaft, Westfälische Wilhelms-Universität Münster, Münster, Deutschland

Philipp Niemann Nationales Institut für Wissenschaftskommunikation gGmbH, Heidelberg, Deutschland

Thomas Quast com.X Institut, Bochum, Deutschland,

Eva Roth Robert Bosch Stiftung GmbH, Stuttgart, Deutschland

Oliver Scheel ZIRIUS Universität Stuttgart, Stuttgart, Deutschland

Andreas M. Scheu Berlin-Brandenburgische Akademie der Wissenschaften, Transfer Unit Wissenschaftskommunikation, Berlin, Deutschland

Yannic Scheuermann Nationales Institut für Wissenschaftskommunikation gGmbH, Heidelberg, Deutschland

Corinna Schuster Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland

Marc Stadtler Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland

Tanja M. Straka Institut für Ökologie, Technische Universität Berlin, Berlin, Deutschland

Vanessa van den Bogaert Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland

Sophia C. Volk Institut für Kommunikationswissenschaft und Medienforschung, Universität Zürich, Zürich, Schweiz

André Weiß Department für Wissenschaftskommunikation, Institut für Technikzukünfte (ITZ), Karlsruher Institut für Technologie (KIT), Karlsruhe, Deutschland

Nina Wicke Institut für Kommunikationswissenschaft der Technischen Universität Braunschweig, Braunschweig, Deutschland

Joachim Wirth Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland

Eva Wollmann ifok GmbH, Düsseldorf, Deutschland

David Ziegler Museum für Naturkunde – Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Deutschland

Ricarda Ziegler Wissenschaft im Dialog, Berlin, Deutschland



Wissenschaftskommunikation evaluieren – mit Methode(n)

Philipp Niemann, Vanessa van den Bogaert und Ricarda Ziegler

Mehr als 20 Jahre nach der Unterzeichnung des Memorandums zu „Public Understanding of Sciences and Humanities" durch die deutschen Wissenschaftsorganisationen hat sich in Deutschland eine große Anzahl an Akteur:innen und Netzwerken der Wissenschaftskommunikation ausdifferenziert, die auf eine Vielzahl von Formaten und Kanälen zurückgreifen kann, um über Wissenschaft und Forschung öffentlich zu kommunizieren. Im Kontext der Diskussion gesellschaftlicher Phänomene wie Fake News oder populistischer Strömungen sowie von großen Herausforderungen wie der Bewältigung des menschengemachten Klimawandels oder der Coronapandemie hat die Wissenschaftskommunikation auch im politischen und öffentlichen Raum zunehmend an Bedeutung gewonnen.

Dabei steht heute nicht mehr die Frage im Mittelpunkt, ob es Aktivitäten und Maßnahmen der gesellschaftlichen Auseinandersetzung mit wissenschaftlichen Inhalten, Forschungsergebnissen und den Kontexten ihrer Entstehung braucht, sondern in den letzten Jahren wurde besonders die Beförderung *guter* Wissenschaftskommunikation und die Sicherung der *Qualität* öffentlicher Kommunikation über Wissenschaft und Forschung thematisiert (Allianz der Wissenschaftsorganisationen 2020; BMBF 2019, 2022; Wissenschaftsrat 2021).

P. Niemann (✉)
Nationales Institut für Wissenschaftskommunikation gGmbH, Heidelberg, Deutschland
E-Mail: niemann@nawik.de

V. van den Bogaert
Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland
E-Mail: vanessa.vandenbogaert@ruhr-uni-bochum.de

R. Ziegler
Wissenschaft im Dialog, Berlin, Deutschland
E-Mail: ricarda.ziegler@w-i-d.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,* https://doi.org/10.1007/978-3-658-39582-7_1

Gleichzeitig gibt es eine große Anzahl an Akteur:innen der Wissenschaftskommunikation – mit unterschiedlichen Funktionen, Rollenzuschreibungen und Selbstverständnissen –, die Erfahrungen aus der Praxis reflektieren, Best Practice und professionelle Expertise teilen (WiD und BVHK 2016). Neben den Erkenntnissen der *science of science communication* – ein Forschungsfeld, das sich international und zunehmend auch in Deutschland konsolidiert (Bonfadelli et al. 2017; Guenther und Joubert 2017; Rauchfleisch und Schäfer 2018) – bieten auch Ergebnisse von (Selbst-)Evaluationen wissenschaftskommunikativer Maßnahmen die Möglichkeit, durch einen systematischen Prozess und den Einsatz passender Methoden Aussagen über die Zielerreichung, die Effektivität und die Qualität von Wissenschaftskommunikation zu treffen und diese damit zukünftig informierter und besser zu gestalten (Jensen und Gerber 2020; Nisbet und Scheufele 2009; Scheufele 2022).

International wird dies anhand der Konzepte einer *effective, strategic* (Besley 2020) oder *evidence-based* (Jensen und Gerber 2020) *science communication* und der *evaluation* und des *assessments* ihres *impacts* diskutiert (Fischhoff 2018; Pellegrini 2021 oder Spicer 2017).

Auch in Deutschland wurden und werden immer wieder einzelne Projekte und Formate der Wissenschaftskommunikation evaluiert. Gleichzeitig gibt es noch keine weitverbreitete, selbstverständliche Evaluationskultur in der deutschen Wissenschaftskommunikationslandschaft (Ziegler und Hedder 2020; Ziegler et al. 2021) – wobei dies auch international noch nicht der Fall ist (Jensen 2014).

Im Zuge der Bereitstellung zunehmender Ressourcen und neuer Fördermöglichkeiten für Wissenschaftskommunikation in Deutschland werden auch Fragen nach ihrer Wirkung, Zielerreichung und Effektivität relevanter, und Akteur:innen sind vielfach angehalten, ihre wissenschaftskommunikativen Aktivitäten und Maßnahmen evaluativ oder durch Forschung zu begleiten. Evaluationsbemühungen in der Wissenschaftskommunikation sollen durch diesen Band unterstützt werden, der daher – und auch wegen der verstärkten Thematisierung von Evaluationen in der internationalen Wissenschaftskommunikationslandschaft – Informationen zu den verschiedenen Evaluationsmethoden explizit auf Deutsch zur Verfügung stellt, sowie Beispiele vorstellt, die Orientierung für die Wissenschaftskommunikationspraxis in Deutschland und im deutschsprachigen Raum geben.

Der Band fokussiert dabei die Methoden, mit denen qualitativ hochwertige, wissenschaftlich valide und systematisch vergleichbare Evaluationen von wissenschaftskommunikativen Maßnahmen möglich werden – ergänzt um einige Verfahren, die in der evaluativen Praxis von besonderer Relevanz sind. Er ist an der Schnittstelle zwischen der *science of science communication* und der evaluativen

Praxis angesiedelt und adressiert einerseits Fachwissenschaftler:innen etwa der Wissenschaftskommunikation, der Medien- oder der Kommunikationswissenschaft bzw. den einschlägigen Lehrbetrieb an Universitäten und Hochschulen. Andererseits werden auch Kommunikator:innen und Wissenschaftler:innen aus anderen Fachdisziplinen angesprochen, die im Zusammenhang mit aktiver Wissenschaftskommunikation auch mit deren Evaluation befasst sind oder es in Zukunft sein werden.

Dieser Mehrfachadressierung versucht der Band durch seine spezifische Struktur Rechnung zu tragen: Nach einem Einführungsteil zu Grundlagen und Perspektiven auf Evaluation stehen in Forschung und Praxis besonders relevante Einzelmethoden bzw. Methodengruppen im Fokus des Bandes. Diese werden – wo immer möglich – in zwei Beiträgen mit unterschiedlichem Schwerpunkt beleuchtet. Während ein tendenziell stärker fachwissenschaftlicher Beitrag zunächst die Grundlagen der jeweiligen Methode oder Methodengruppe in den Blick nimmt, widmet sich der daran anschließende Praxisbeitrag einem Anwendungsfall aus der Evaluationspraxis der Wissenschaftskommunikation, bei dem die jeweilige Methode(ngruppe) zum Einsatz kam. Damit eignet sich der Band sowohl zur Vertiefung und Erweiterung der eigenen Methodenkenntnisse auf aktuellem fachwissenschaftlichen Stand als auch als konkrete Handreichung im praktischen Evaluationsalltag, etwa als Wissenschaftskommunikator:in.

Dem Band liegt ein weites Verständnis von Wissenschaftskommunikation zugrunde, wie es von Schäfer, Kristiansen und Bonfadelli (2015) vorgeschlagen wird. Danach werden „alle Formen, von auf wissenschaftliches Wissen oder wissenschaftliche Arbeit fokussierter Kommunikation, sowohl innerhalb als auch außerhalb der institutionalisierten Wissenschaft, inklusive ihrer Produktion, Inhalte, Nutzung und Wirkungen“ (S. 13) als Wissenschaftskommunikation verstanden. Der Bezugsrahmen der Autor:innen ist der Bereich der externen Wissenschaftskommunikation, verstanden als eine Wissenschaftskommunikation, die sich aus der Wissenschaft heraus an einen Personenkreis außerhalb der eigenen fachwissenschaftlichen Peers richtet. Dies bedeutet jedoch in keiner Weise, dass sich die vorgestellten Evaluationsmethoden nicht auch zur Anwendung auf Maßnahmen aus dem wissenschaftsinternen Bereich oder dem Bereich des Wissenschaftsjournalismus eigneten.

Der Begriff *Evaluation* findet sich sowohl im wissenschaftlichen als auch im alltäglichen Sprachgebrauch. Aufgrund der Vielfalt der Begriffsverwendung von Evaluationen, ihrer Anwendungsbereiche, Aufgaben und zugrunde liegenden Denk- und Handlungskonzepten sowie der zunehmenden alltagssprachlichen Verwendung, mag der Eindruck entstehen, dass es genauso viele Definitionen

des Begriffs wie Evaluator:innen gibt (Kromrey 2001; Wottawa und Thieraus 1998). Während annähernd alles evaluiert werden kann und damit die Liste an möglichen Evaluationsgegenständen unerschöpflich erscheint (Balzer und Beywl 2018), werden gleichzeitig unterschiedliche Begrifflichkeiten für ein und denselben Bedeutungszusammenhang genutzt.[1] Die ausgeprägte Heterogenität dieses Feldes spiegelt sich auch in den unterschiedlichen disziplinären Ansätzen und Zugängen und damit in einer großen methodischen Vielfalt wider – wobei Evaluationen dabei immer ein besonders starker Kontextbezug auszeichnet. Diese Tatsache mag auch ursächlich dafür sein, dass eine Verständigung auf eine allumfassende Definition im Wortsinne nicht zielführend erscheint (von Werthern 2020).

Für diesen Band schließen wir uns der Idee einer sogenannten Arbeitsdefinition an (Balzer und Beywl 2018, S. 17) und verstehen Evaluationen als eine professionelle Forschungspraxis, in welcher durch vielfältige forschungsmethodische Zugänge systematisch Informationen über jeweils sehr spezifische Evaluationsgegenstände gesammelt und anschließend kriteriengeleitet bewertet werden. Sowohl im wissenschaftlichen Diskurs als auch in der Praxisbeschreibung können Evaluationen auf einem Kontinuum zwischen theorie(weiter-)entwickelnder und angewandter Forschung verortet werden. In diesem Band nutzen wir die spezifischen Merkmale: *Ziel und Zweck*, *Herkunft der Fragestellung, Verwendung von Methoden* sowie *Anspruch an die erzielten Ergebnisse,* um einerseits die Spannweite zwischen den Polen aufzuzeigen und andererseits der Arbeitsdefinition eine klare Struktur zu verleihen. Dazu werden im Folgenden Evaluations- und Forschungsvorhaben zunächst konzeptionell nahe den gegenüberliegenden Polen verortet und mit der Einordnung von Begleitforschungsvorhaben beispielhaft aufgezeigt, dass zwischen den zwei Polen – je nach Ausgestaltung der Vorhaben – viel Spielraum zur individuellen Umsetzung besteht.

Evaluationen stellen einen bewussten, (ggf. projektbegleitenden) Forschungs- und Lernprozess dar, in dem konkrete Aktivitäten innerhalb oder die Gesamtheit aller Aktivitäten eines Projekts über eine vorab festgelegte Methodik systematisch betrachtet und bewertet werden. Hierbei können die Inhalte und

[1] An dieser Stelle sei bereits darauf hingewiesen, dass verschiedene Autor:innen (bspw. Döring und Bortz 2016 oder Stockmann und Meyer 2017) den Begriff der Evaluationsforschung durchaus als Synonym für Evaluationen nutzen. In diesem Beitrag repräsentiert dieser Terminus zunächst ausschließlich die Forschung über Evaluationen und damit über die Praxis, die Methoden, aber auch die Nutzung von Evaluationen.

Rahmenbedingungen der Projektaktivitäten, aber auch ihre Leistungen, Produkte und Wirkungen im Fokus stehen. Evaluationen können die Reflexion unterstützen und dabei helfen, Erwartungen zu formulieren oder einen Abgleich mit zuvor gesteckten Zielen ermöglichen, um Erfolge und Misserfolge zu ergründen sowie den Wert der Projektaktivitäten zu erkennen. Dazu werden im Rahmen von Evaluationsvorhaben Daten, Schlussfolgerungen und Bewertungen bereitgestellt, die dem Informationsbedarf der jeweiligen Stakeholder:innen (dies schließt Auftraggeber:innen sowie Beteiligte und Betroffene gleichermaßen ein) gerecht werden müssen. Die im Rahmen einer Evaluation gewonnenen Daten und daraus abgeleiteten Schlussfolgerungen und Bewertungen können für unterschiedliche Zwecke genutzt werden, u. a. als Grundlage für Entscheidungen, zur Rechenschaftslegung oder zur Identifizierung von Optimierungsbedarfen. Evaluationen sind immer empfänger:innenorientiert, und ihre Güte hängt nicht nur von der Zuverlässigkeit und Genauigkeit (Einhaltung von Standards[2]) ab, sondern immer auch von der Akzeptanz der Beteiligten und durch die Betroffenen (siehe dazu auch DeGEval 2017; Döring und Bortz 2016; Stockmann 2010, Stockmann und Meyer 2017; Weiss 1998; Wottawa und Thierau 1998).

Ziel und Zweck von Evaluationen sind immer Bewertungen, die nach explizit formulierten Kriterien erfolgen. Die Fragestellungen von Evaluationen werden durch Auftraggeber:innen oder Anspruchsgruppen oft mit spezifischen inhaltlichen und zeitlichen Vorstellungen, aber auch durch konkrete Budgetierungen gerahmt. Dabei ergibt sich eine Besonderheit für sogenannte Selbstevaluationen, bei denen die praxisgestaltenden Akteur:innen identisch mit den Evaluator:innen sind (DeGEval 2004). Auch bei Selbstevaluationen kann der Auftrag zur Evaluation bspw. von Vorgesetzten, Kolleg:innen oder hierarchisch höheren Instanzen in der Organisation bzw. als Bedingung im Rahmen von Förderrichtlinien erfolgen. Damit wird eine Selbstevaluationsaufgabe den Evaluator:innen angetragen. Sie kann aber auch ohne einen Auftrag im weitesten Sinne gänzlich aus Eigeninitiative vollzogen werden. In beiden Fällen sind die Akteur:innen gleichzeitig praxisverantwortlich sowie evaluationsverantwortlich. Dabei verfolgen Selbstevaluationen meist eine doppelte Zielsetzung: Einerseits geht es um den Gewinn von Informationen und Erkenntnissen, andererseits ist die möglichst unmittelbare Veränderung der Praxis das Ziel (Müller-Kohlenberg und Beywl 2003).

---

[2] Evaluationsstandards aber auch Einhaltung von wissenschaftlichen Gütekriterien.

Zusammenfassend ist das wichtigste Merkmal von Evaluationen, dass diese immer auf einen oder mehrere konkrete, vorab bereits bekannte Zwecke hin durchgeführt werden (Hense 2021). Vier ganz zentrale mögliche Zwecke sind dabei:

1. die Verbesserung bzw. Qualitätssicherung des Gegenstands, bspw. die Passgenauigkeit oder Effektivität von Angeboten für spezifische Zielgruppen,
2. die Rechenschaftslegung, bspw. über ein Programm oder eine Maßnahme, welche aus Drittmitteln finanziert wird,
3. die Unterstützung von Entscheidungen, bspw. über Fortführung oder Einstellung eines Wissenschaftskommunikationsangebots und
4. das Dazulernen, bspw. bei der Erprobung eines innovativen digitalen Dialogportals.

„Mit dieser Zweckgebundenheit teilt Evaluation eine wichtige Gemeinsamkeit mit den Maßnahmen, Projekten oder Praktiken, die sie evaluiert: Sie will grundsätzlich etwas in Bewegung bringen und einen Unterschied in der Praxis machen“, so Hense (2021 S. 3). Evaluationen sind auf dem aufgespannten Kontinuum zwischen theorie(weiter-)entwickelnder und angewandter Forschung somit eher dem Pol des angewandten Forschungstypus zuzuordnen, da sie nicht den originären Zweck haben, einen Beitrag zum aktuellen wissenschaftlichen Diskurs zu leisten (Döring und Bortz 2016), sondern vielmehr auf den konkreten Nutzen für die Praxis hinarbeiten. Der gegenüberliegende Pol der theorie(weiter-)entwickelnden Grundlagenforschung zeichnet sich durch das Ziel aus, möglichst generalisierbare Erkenntnisse bereitzustellen, welche unmittelbar in die Theorie bzw. Forschungsdebatten integriert werden können. Wissenschaftliche Theorien dienen derartiger Forschung bei der Beschreibung, Erklärung und Vorhersage von beobachteten Sachverhalten. Anders als bei Evaluationen müssen sich (Grundlagen-)Forschungsvorhaben nicht primär an außerwissenschaftlichen Verwertungskontexten orientieren. Zwischen Grundlagenforschung und klassischen Evaluationen ist eine dritte Kategorie zu verorten: die (evaluative) Begleitforschung. Während Evaluationen oft sehr kontextspezifisch sind, widmen sich Begleitforschungsvorhaben bspw. der Messung von Effekten in experimentellen Forschungsansätzen, um so Kausalmechanismen mit theoriegeleiteten Forschungsansätzen zu untersuchen. Begleitforscher:innen betrachten systematisch die Veränderungen und die Unterschiede (meist Outcome, seltener Impact), die ursächlich auf bestimmte Maßnahmen zurückgeführt werden können. Gleichzeitig kann an Begleitforschungsvorhaben der Anspruch gestellt werden, die Relevanz der Erkenntnisse auch für ausgewählte Praxisfelder zu

antizipieren und den Erkenntnistransfer in die Praxis mitzuplanen sowie die Zusammenarbeit mit Praktiker:innen zu intensivieren.

In den Praxisbeiträgen dieses Bandes finden sich Evaluationsbeispiele, die auf unterschiedlichen Punkten des Kontinuums zwischen Grundlagenforschung und angewandter Forschung zu verorten sind. Auch unterscheiden sich die Praxisbeiträge darin, ob sie direkte Outputs von Wissenschaftskommunikationsprojekten wie Besucher:innenzahlen oder Abrufe von bereitgestellten Informationsmaterialien erhoben haben, oder ob versucht wurde, die tatsächlichen individuellen oder gar gesellschaftlichen Wirkungen von Wissenschaftskommunikation zu erfassen. Beide Evaluationsansätze sollen hier in ihrer jeweiligen Bedeutung anerkannt werden. Gerade im Kontext von (Selbst-) Evaluationen in der Praxis können wertvolle Erkenntnisse für die Überprüfung und Weiterentwicklung von Wissenschaftskommunikation generiert werden, wenn aussagekräftige Daten auf einer Outputebene erhoben werden, statt mit zu geringen Ressourcen oder Kenntnissen zu versuchen, eine Wirkung zumal auf gesellschaftlicher Ebene nachzuweisen (vgl. hierzu auch King et al. 2015). Andererseits ist zu betonen, dass insgesamt eine stärkere Wirkungsorientierung und entsprechend darauf ausgerichtete Evaluationen in der Wissenschaftskommunikation zu begrüßen sind, so wie dies auch von der internationalen Forschungscommunity gefordert wird (vgl. hierzu auch Jensen 2015; Pellegrini 2021; Weitkamp 2015).

Konkret sind folgende Inhalte in diesem Band zusammengetragen:

Ricarda Ziegler, Imke Hedder und Liliann Fischer beleuchten zu Beginn unter Einbezug verschiedener Datenquellen, wie Wissenschaftskommunikation derzeit in Deutschland evaluiert wird, welchen Schwierigkeiten dies momentan unterliegt und zeigen notwendige Veränderungen für eine aussagekräftige Evaluationspraxis auf.

Sophia Volk verortet anschließend die Evaluation von Wissenschaftskommunikation in der (klassischen) Evaluationsforschung und führt in Logiken und Modelle von Evaluationen sowie sozial- und betriebswirtschaftliche Methoden für Evaluationen ein.

Vanessa van den Bogaert stellt in ihrem Beitrag die fachlichen Standards für Evaluation der Gesellschaft für Evaluation *(DeGEval)* vor. Dazu beleuchtet sie die vier Qualitätsdimensionen von Evaluationen (Nützlichkeit, Durchführbarkeit, Fairness, Genauigkeit) und zeigt auf, dass die Verantwortung für qualitativ hochwertige Evaluationen gemeinsam getragen werden kann.

Christoph Böhmert und Ferdinand Abacioglu eröffnen den Hauptteil des Bandes zu einzelnen Evaluationsmethoden der Wissenschaftskommunikation mit einem Grundlagenbeitrag zur Methode der quantitativen Befragung. Valerie

Knapp und Vanessa van den Bogaert schließen daran mit ihrem Praxisbeitrag zur gleichen Methode an. Anwendungsbeispiel ist eine mehrwellige, onlinegestützte Panelstudie, die begleitend zu einer internationalen Citizen-Science-Aktion im Rahmen einer Begleitforschung durchgeführt wurde.

Julia Metag und Andreas Scheu steuern den Grundlagenbeitrag zur Methode der qualitativen Befragung bei. Im nachfolgenden Praxisbeitrag von Imke Hedder, Ricarda Ziegler, Bonnie Dietermann und David Ziegler steht die Anwendung von leitfadengestützten Vorher-Nachher-Interviews im Rahmen der Evaluation des Wissenschaftsvarietés *Glitzern & Denken* im Mittelpunkt.

André Weiß widmet sich der Methode der Beobachtung, wobei er neben der Auseinandersetzung mit grundlegenden Kriterien und Formen der Methode auch eine Reihe von Anwendungsszenarien und Beispielen für Beobachtungen im Rahmen von evaluativer Wissenschaftskommunikationsforschung behandelt. Der Text vereint damit Grundlagen- und Praxisbeitrag.

Armin Hempel befasst sich in seinem Grundlagenbeitrag mit Methoden zur Nutzungsdatenanalyse digitaler Medien. Till Bruckermann und Hannah Greving liefern den dazugehörigen Praxisbeitrag und fokussieren darin am Beispiel eines Bürgerwissenschaftsprojekts zu Wildtieren die Analyse von Nutzungsdaten der Teilnehmenden zur Häufigkeit und Art der Beteiligung an digitalen Projektaktivitäten und die daraus ableitbaren Implikationen.

Philipp Niemann und Yannic Scheuermann setzen sich in ihrem Grundlagenbeitrag mit verschiedenen physiologischen Messmethoden auseinander. Der anschließende Praxisbeitrag von Christian Humm und Philipp Niemann stellt mit der Methode der Blickaufzeichnung eines dieser Verfahren in den Mittelpunkt – am Beispiel eines evaluativen Forschungsprojekts zu *KATRIN VR,* einer Virtual-Reality-Umgebung aus dem Themenbereich der Physik.

Sabrina Kessler und Nina Wicke haben den Grundlagenbeitrag zur Inhalts- und Medienanalyse verfasst. Der Praxisbeitrag von Rüdiger Goldschmidt und Oliver Scheel gibt Einblick in eine umfangreiche Studie zum Vergleich von Dialog- und Beteiligungsformaten und demonstriert die Rolle und Leistungen der Methode der Inhaltsanalyse in diesem Kontext.

Joachim Wirth und Jens Fleischer geben in ihrem Grundlagenbeitrag einen Einblick in die Entwicklung und Interpretation von Testverfahren. Passend dazu stellen Till Bruckermann, Tanja Straka und Moritz Krell im dazugehörigen Praxisbeitrag eine Vorlage für einen Test über Fähigkeiten zum wissenschaftlichen Denken bereit.

Marc Stadtler und Corinna Schuster führen in ihrem Grundlagenbeitrag in Experimentaldesigns für die Evaluation von Maßnahmen der Wissenschaftskommunikation ein, und Hannah Greving, Till Bruckermann und Joachim

Kimmerle veranschaulichen diese anschließend an Praxisbeispielen im Kontext von Bürgerwissenschaftsprojekten.

Anschließend berichten Eva Wollmann und Jacob Birkenhäger aus der Praxis von Beteiligungsverfahren und zeigen Möglichkeiten und Methoden auf, mit denen im konkreten Kontext von Veranstaltungen und Workshops Feedback zu Evaluationszwecken eingeholt werden kann.

Den Abschluss des Bandes bildet die Vorstellung der Evaluation der Dialogveranstaltung *Mensch Wissenschaft!* von Markus Gabriel, Isabella Kessel, Thomas Quast und Eva Roth. Zentral ist hierbei die Anwendung eines Multimethodenansatzes, in welchem dem Erkenntnisinteresse innerhalb der Rahmenbedingungen der Evaluation durch unterschiedliche Verfahren nachgegangen wird.

Im Zentrum dieses Bandes stehen Beiträge, die sich aus einer methodisch orientierten Sichtweise darauf konzentrieren, Evaluationen in ihrem Potenzial sichtbar werden zu lassen und damit auch konkrete Beispiele für das Evaluieren unterschiedlicher – auch informeller oder interaktiver – Formate und Aktivitäten von Wissenschaftskommunikation (vgl. Grand und Sardo 2017) mit unterschiedlichen Evaluationsfragen zu unterschiedlichen Zeitpunkten im Prozess der Wissenschaftskommunikation (Pellegrini 2021) präsentieren. Es wird deutlich, dass das Postulat der Vergleichbarkeit nicht zum einzigen Ziel von Evaluationen werden sollte. Vielmehr bestehen konstruktive Evaluationen aus überzeugend dargestellten Kernleistungen, die die Individualität eines Programms bzw. einer Maßnahme nachvollziehbar sichtbar machen.

In vielen Beiträgen, die ein konkretes Praxisbeispiel vorstellen, stehen Bürger:innen, Besucher:innen, Schüler:innen, Lai:innen o. Ä. und ihre Wahrnehmung einer Aktivität oder die Wirkung eines Projekts auf sie im Fokus der Evaluationen. Dies soll nicht darüber hinwegtäuschen, dass auch die Wirkungen von öffentlicher Wissenschaftskommunikation auf beteiligte Forschende oder die (ggf. auch unintendierten) Rückwirkungen auf eine Forschungseinrichtung oder ins Wissenschaftssystem Gegenstand und damit Objekt der Evaluationen von Wissenschaftskommunikation sein können.

Mit diesem multiperspektivischen Band möchten die Herausgeber:innen einen Beitrag zu einer aussagekräftigen und vielfältigen Evaluationspraxis in der deutschsprachigen Wissenschaftskommunikationscommunity an der Schnittstelle von Praxis und Forschung und gemeinsam mit Akteur:innen aus beiden Bereichen leisten. Die Hoffnung ist, Evaluation zu einem (selbstverständlichen) Lernprozess für die Community der Wissenschaftskommunikation zu machen und basierend auf ihren Ergebnissen Akteur:innen der Wissenschaftskommunikation dazu zu inspirieren, über ihre Ziele und Aktivitäten zu

reflektieren, um langfristig gute und wirkungsvolle Wissenschaftskommunikation für und mit der Öffentlichkeit zu gestalten.

Der Band versucht gezielt, das Thema der Evaluation von Wissenschaftskommunikation sowohl durch Autor:innen aus der Praxis als auch durch Wissenschaftler:innen unterschiedlicher Fachkulturen zu beleuchten. Damit leistet er einen Beitrag zum dringend notwendigen Ausbau der Vernetzung zwischen Wissenschaft und Praxis der Wissenschaftskommunikation (Jensen und Gerber 2020; Han und Stenhouse 2022; Scheufele 2022).

Der besondere Dank der Herausgeber:innen gilt allen Autor:innen, die mit viel Engagement und Ausdauer das Zustandekommen dieses Bandes ermöglicht haben. Eine ebenso herzlicher Dank gilt darüber hinaus dem Bundesministerium für Bildung und Forschung und der Klaus Tschira Stiftung, die es durch die Übernahme der Open-Access-Veröffentlichungskosten ermöglicht haben, dass dieser Band, der sich explizit an Akteur:innen aus Praxis und Forschung der Wissenschaftskommunikation richtet, nicht nur einem wissenschaftlichen Publikum an Forschungs- und Bildungseinrichtungen zugänglich ist, sondern von vielen eingesehen und herangezogen werden kann.

Literatur

Allianz der Wissenschaftsorganisationen (2020) 10-Punkte-Plan zur Wissenschaftskommunikation. https://www.allianz-der-wissenschaftsorganisationen.de/themen-stellungnahmen/10-punkte-plan-zur-wissenschaftskommunikation/. Zugegriffen: 29. Aug. 2022

Balzer L, Beywl W (2018) Evaluiert – erweitertes Planungsbuch für Evaluationen im Bildungsbereich (2., überarbeitete Aufl.). hep verlag, Bern

Besley JC (2020) Five thoughts about improving science communication as an organizational activity. J Commun Manag 24(3):155–161. https://doi.org/10.1108/JCOM-03-2020-0022

Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (2017) Das Forschungsfeld Wissenschaftskommunikation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer M (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-12898-2_1

BMBF (2019) Grundsatzpapier des Bundesministeriums für Bildung und Forschung zur Wissenschaftskommunikation. https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/1/24784_Grundsatzpapier_zur_Wissenschaftskommunikation.pdf?__blob=publicationFile&v=4. Zugegriffen: 29. Aug. 2022

BMBF (2022) #FactoryWisskomm: Handlungsperspektiven für die Wissenschaftskommunikation. https://www.bmbf.de/bmbf/shareddocs/downloads/files/factorywisskommpublikation.pdf. Zugegriffen: 29. Aug. 2022

DeGEval [Gesellschaft für Evaluation] (2004) Empfehlungen zur Anwendung von Standards für Evaluation im Handlungsfeld der Selbstevaluation. DeGEval – Gesellschaft für Evaluation, Alfter
DeGEval [Gesellschaft für Evaluation] (2017) Standards für Evaluation Erste Revision 2016. DeGEval – Gesellschaft für Evaluation, Mainz
Döring N, Bortz J (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer, Berlin
Fischhoff B (2018) Evaluating science communication. Proc Natl Acad Sci 116(16):7670–7675. https://doi.org/10.1073/pnas.1805863115
Grand A, Sardo AM (2017) What works in the field? Evaluating informal science events. Front Commun 2(22). https://doi.org/10.3389/fcomm.2017.00022
Guenther L, Joubert M (2017) Science communication as a field of research: identifying trends, challenges and gaps by analysing research papers. JCOM 16(02):A02. https://doi.org/10.22323/2.16020202
Jensen E (2014) The problems with science communication evaluation. JCOM 13(01):C04
Jensen E (2015) Highlighting the value of impact evaluation: enhancing informal science learning and public engagement theory and practice. JCOM 14(03):Y05
Jensen E, Gerber A (2020) Evidence-based science communication. Front Commun 4:78. https://doi.org/10.3389/fcomm.2019.00078
Han H, Stenhouse N (2022) Bridging the research-practice gap in climate communication: lessons from one academic-practitioner collaboration. Sci Commun 37(3):396–404. https://doi.org/10.1177/1075547014560828
Hense JU (2021) Nicht jede Evaluation ist eine gute Evaluation. Warum gute Evaluationen fachliche Standards berücksichtigen sollten, FORUM Sexualaufklärung und Familienplanung: Informationsdienst der Bundeszentrale für gesundheitliche Aufklärung (BZgA), 1, S 3–5
King H, Steiner K, Hobson M, Robinson A, Clipson H (2015) Highlighting the value of evidence-based evaluation: pushing back on demands for 'impact'. J Sci Commun 14(02). https://doi.org/10.22323/2.14020202
Kromrey H (2001) Evaluation von Lehre und Studium – Anforderungen an Methodik und Design. http://www.hkromrey.de/eval_kromrey_in-Spiel.pdf. Zugegriffen: 22. July 2022
Müller-Kohlenberg H, Beywl W (2003) Standards der Selbstevaluation – Begründung und aktueller Diskussionsstand. Z Eval 1:65–75
Nisbet MC, Scheufele DA (2009) What's next for science communication? Promising directions and lingering distractions. Am J Bot 96(10):1767–1778. https://doi.org/10.3732/ajb.0900041
Pellegrini G (2021) Evaluating science communication: concepts and tools for realistic assessment. https://doi.org/10.4324/9781003039242-17-16
Rauchfleisch A, Schäfer MS (2018) Structure and development of science communication research: co-citation analysis of a developing field. JCOM 17(03):A07. https://doi.org/10.22323/2.17030207
Schäfer M, Kristiansen S, Bonfadelli H (2015) Wissenschaftskommunikation im Wandel: Relevanz, Entwicklung und Herausforderungen des Forschungsfeldes. In: Schäfer M, Kristiansen S, Bonfadelli H (Hrsg) Wissenschaftskommunikation im Wandel. Herbert von Halem, Köln, S 10–42

Scheufele DA (2022) Thirty years of science–society interfaces: what's next? Public Underst Sci 31(3):297–304. https://doi.org/10.1177/09636625221075947
Spicer S (2017) The nuts and bolts of evaluating science communication activities. Semin Cell Dev Biol 70:17–25. https://doi.org/10.1016/j.semcdb.2017.08.026
Stockmann, R (2010) Rolle der Evaluation in der Gesellschaft. In: Stockmann R, Meyer, W (Hrsg) Evaluation. Eine Einführung. Budrich, Opladen, S 15–54
Stockmann R, Meyer W (2017) Einleitung. In: Stockmann R, Meyer W (Hrsg) Die Zukunft der Evaluation. Trends, Herausforderungen, Perspektiven, Bd 13. Waxmann, Münster, S 9–20
von Werthern, A. (2020) Theoriebasierte Evaluation – Entwicklung und Anwendung eines Verfahrensmodells zur Programmtheoriekonstruktion, Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-27579-2
Weitkamp E (2015) Between ambition and evidence. JCOM 14(02)
Weiss CH (1998) Evaluation. Methods for Studying Programs and Policies, 2. Ausgabe. Prentice Hall, New Jersey
Wottawa H, Thierau H (1998) Lehrbuch Evaluation. Huber, Bern
Wissenschaft im Dialog und Bundesverband Hochschulkommunikation [WiD und BVHK] (2016) Leitlinien zur guten Wissenschafts-PR. https://www.wissenschaft-im-dialog.de/fileadmin/user_upload/Ueber_uns/Gut_Siggen/Dokumente/Leitlinien_zur_guten_Wissenschafts-PR.pdf. Zugegriffen: 29. Aug. 2022
Wissenschaftsrat (2021) Wissenschaftskommunikation Positionspapier. https://www.wissenschaftsrat.de/download/2021/9367-21.pdf?__blob=publicationFile&v=10. Zugegriffen: 29. Aug. 2022
Ziegler R, Hedder IR (2020) Evaluationspraktiken in der Wissenschaftskommunikation – Eine Betrachtung veröffentlichter Evaluationsberichte im deutschsprachigem Raum. Wissenschaft im Dialog, Berlin
Ziegler R, Hedder IR, Fischer L (2021) Evaluation of science communication: current practices, challenges, and future implications. Front Commun 6:669744. https://doi.org/10.3389/fcomm.2021.669744

Philipp Niemann ist stellvertretender Direktor und wissenschaftlicher Leiter des Nationalen Instituts für Wissenschaftskommunikation (NaWik). Zuvor war er als Nachwuchsgruppenleiter im Department für Wissenschaftskommunikation am Karlsruher Institut für Technologie (KIT) tätig. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftskommunikation, qualitative Rezeptionsforschung und politische Kommunikation.

Vanessa van den Bogaert ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Lehr-Lernforschung der Ruhr-Universität Bochum. Sie widmet sich in ihren Forschungsschwerpunkten der wissenschaftlichen Begleitforschung von Citizen-Science-Projekten sowie der Grundlagenforschung zur Interessengenese an außerschulischen Lernorten. Sie leitet die Arbeitsgruppe Science of Citizen *Science* in Zusammenarbeit mit *Bürger schaffen Wissen*.

Ricarda Ziegler ist Leiterin des Bereichs Qualität & Transfer bei Wissenschaft im Dialog (WiD) – der deutschen Organisation für Wissenschaftskommunikation. Sie verantwortet dort u. a. die Impact Unit, die sich Fragen der Wirkung und Evaluation von Wissenschaftskommunikation widmet. Außerdem leitet sie das bevölkerungsrepräsentative Wissenschaftssurvey Wissenschaftsbarometer. Ricarda Ziegler hat einen Hintergrund in der Politikwissenschaft.

Grundlagen der Evaluation von Wissenschaftskommunikation



Herausforderungen der aktuellen Evaluationspraxis in der Wissenschaftskommunikation in Deutschland

Ricarda Ziegler, Imke Hedder und Liliann Fischer

Zusammenfassung

Evaluationen bieten einen wichtigen Mehrwert für Wissenschaftskommunikation, denn anhand ihrer Ergebnisse lässt diese sich zukünftig zielorientiert und effektiv gestalten. Zur Zeit steht die Evaluation von Wissenschaftskommunikation in Deutschland allerdings noch vor Herausforderungen. So ergeben sich bereits vor Beginn der Evaluationen Probleme durch fehlende strategische Planung von Wissenschaftskommunikation. Darüber hinaus mangelt es bei Evaluationen oft an passenden Evaluationsdesigns und geeigneten Datenerhebungsmethoden. Zu guter Letzt erschwert das in der deutschen Wissenschaftskommunikationspraxis vorherrschende Bild von Evaluation einen kollektiven und konstruktiven Lernprozess für die Wissenschaftskommunikation. Diese Herausforderungen gilt es zu überwinden, damit Evaluation als kollektiver Reflexionsprozess zur konstruktiven Weiterentwicklung von Wissenschaftskommunikation beitragen kann.

R. Ziegler (✉) · I. Hedder · L. Fischer
Wissenschaft im Dialog, Berlin, Deutschland
E-Mail: ricarda.ziegler@w-i-d.de

I. Hedder
E-Mail: imke.hedder@w-i-d.de

L. Fischer
E-Mail: liliann.fischer@w-i-d.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_2

1 Wie lässt sich die Evaluationspraxis in der Wissenschaftskommunikation beschreiben?

Gemeinsam mit Erkenntnissen aus der Forschung zu Wissenschaftskommunikation stellen Einblicke aus Evaluationen eine wichtige Grundlage für die Wissenschaftskommunikationspraxis dar. Basierend auf ihren Ergebnissen können Aussagen über die Effekte und Wirkungen von Wissenschaftskommunikation bei den erreichten Zielgruppen getroffen und Aktivitäten insgesamt strategisch und damit wirkungsvoll und ziel(gruppen)orientiert gestaltet werden (Pellegrini 2021; siehe auch van den Bogaert in diesem Band). Einerseits messen inzwischen auch viele in Deutschland Evaluationen in der Wissenschaftskommunikation eine größere Bedeutung bei (siehe bspw. BMBF 2019[1]). Andererseits scheint der Weg zu einer aussagekräftigen Evaluationspraxis basierend auf methodisch sinnvoll gestalteten Evaluationsdesigns noch weit. Den aktuellen Status der Evaluationspraxis der Wissenschaftskommunikation[2] in Deutschland und damit verbundene Herausforderungen schildern wir in diesem Beitrag.

Dem Spannungsverhältnis zwischen zunehmender Relevanz und mangelnder Qualität sowie Aussagekraft von Evaluationen scheint sich die Community der Praktiker:innen der Wissenschaftskommunikation durchaus bewusst: In einer Ende 2019 von Wissenschaft im Dialog (WiD) durchgeführten Online-Umfrage unter 109 deutschen Wissenschaftskommunikator:innen (im Folgenden Community-Befragung) stimmte zwar eine Mehrheit zu, dass Evaluationen in der Wissenschaftskommunikation wichtig (87 %) und deren Ergebnisse meist erkenntnisreich seien (58 %, n = 82, Impact Unit 2019, S. 28), es fühlten sich aber nur 37 % in der Lage, die Qualität von Evaluationen in der Wissenschaftskommunikation zu beurteilen. Bei der Einschätzung der eigenen Fähigkeiten zur

[1] Auch die Veröffentlichung dieses Bands unter Mitwirkung einer großen Bandbreite verschiedener Forschender und Praktiker:innen verdeutlicht dies.

[2] Hierbei wird unter Wissenschaftskommunikation in diesem Beitrag die externe Wissenschaftskommunikation verstanden (vgl. Schäfer et al. 2015) und der Fokus der in diesem Beitrag präsentierten Analysen bezieht sich auf die institutionelle oder eigenvermittelte Wissenschaftskommunikation – also die Wissenschaftskommunikation aus dem Wissenschaftssystem heraus an außerwissenschaftliche Öffentlichkeiten, beispielsweise durch Universitäten oder Forschungsinstitute. Dies geschieht, um den Analysegegenstand einzugrenzen und um möglichst präzise Aussagen zuzulassen, und soll keineswegs die gesellschaftliche Bedeutung von Wissenschaftskommunikation in anderen sozialen Kontexten (zum Beispiel den Wissenschaftsjournalismus) geringschätzen.

Umsetzung guter Evaluationen ergibt sich ebenfalls ein geteiltes Bild (n = 82, ebd., S. 29).

Derartige Surveyergebnisse in Form von Selbsteinschätzungen der Praktiker:innen sind nur eine Möglichkeit, Informationen über die Evaluationspraxis in der Wissenschaftskommunikation zu erlangen. In Ermangelung einer systematischen, umfassenden Analyse der Arten und Weisen, wie Wissenschaftskommunikation in Deutschland evaluiert wird, zieht der folgende Beitrag verschiedene Datenquellen heran. Neben den Ergebnissen der Community-Befragung wird auf die Ergebnisse einer Analyse von 51 veröffentlichten Evaluationsberichten[3] deutschsprachiger Wissenschaftskommunikationsprojekte zurückgegriffen. Diese Analyse wurde im Rahmen des BMBF-geförderten Projekts „Impact Unit – Wirkung und Evaluation in der Wissenschaftskommunikation" im Jahr 2020 bei WiD durchgeführt und baut auf einer Stichwortrecherche öffentlich zugänglicher Evaluationsberichte aus den Jahren 2000 bis 2019 auf (Ziegler und Hedder 2020). Ergänzt wird dies durch die Ergebnisse verschiedener Diskussionsrunden und Workshops mit Stakeholder:innen der Wissenschaftskommunikationspraxis und -forschung sowie aus Wissenschaftsmanagement und -politik, die im Rahmen des Projekts Impact Unit in den Jahren 2020 und 2021 durchgeführt wurden.

Keine dieser Datenquellen erlaubt repräsentative Einblicke in die Evaluationspraxis der Wissenschaftskommunikation in Deutschland. Dies bedürfte eines sehr hohen Aufwands für die inzwischen stark ausdifferenzierte Branche der Wissenschaftskommunikation (Bonfadelli et al. 2017) und ihrer Evaluationen. Evaluationen werden nicht nur von Wissenschaftskommunikator:innen selbst durchgeführt, sondern finden auch in Form von wissenschaftlicher Begleitforschung oder als Auftragsvergaben an externe Evaluationsdienstleistende statt. Die in diesem Beitrag geschilderten Ausführungen, die sich besonders auf die von Wissenschaftskommunikator:innen gestalteten (Selbst-)Evaluationen beziehen, sind daher als Annäherung an die Evaluationspraxis zu verstehen und sollen einen Mehrwert durch die Zusammenführung der Ergebnisse und Erkenntnisse aus den verschiedenen vorliegenden Quellen liefern.[4]

---

[3] Insgesamt lagen 51 Dokumente zur näheren Untersuchung vor, in denen 55 evaluierte Projekte beschrieben werden. Sofern nicht explizit vermerkt, bezieht sich die Analyse im Folgenden auf jene Projekte und damit auf 55 Fälle.

[4] Eine Zusammenführung der Ergebnisse mit Fokus auf einer internationalen Kontextualisierung haben die Autorinnen bereits im Journal „Frontiers in Communication – Science and Environmental Communication" veröffentlicht (Ziegler et al. 2021).

Die in den nächsten Abschnitten geschilderten Herausforderungen bei der Durchführung von Evaluationen in der Wissenschaftskommunikation in Deutschland fügen sich dabei in bisherige Betrachtungen des Status Quo der internationalen Evaluationspraxis ein. So wird beispielsweise in der Forschung zu Wissenschaftskommunikation unter verschiedenen Bezeichnungen (z. B. strategic, effective, evidence-based, more quality) gefordert, Wissenschaftskommunikation stärker ziel- und wirkungsorientiert zu betreiben (Besley et al. 2019; Scheufele et al. 2021; Dudo und Besley 2016), über ihre Qualität zu reflektieren (Mannino et al. 2021; Olesk et al. 2021; Wicke 2021) und dafür Forschungserkenntnisse und Evaluationsergebnisse zu nutzen (Fischhoff 2019; National Academies of Sciences, Engineering, and Medicine 2017; Pellegrini 2021). Gleichzeitig wird auch erkannt, welche Herausforderungen in der Praxis die Umsetzung eines derartigen Ansatzes und einer damit aussagekräftigen Evaluationspraxis bisher erschweren (Jensen 2020; Jensen und Gerber 2020; Peterman et al. 2020). Jensen merkt bereits 2014 an, dass Evaluationen von Wissenschaftskommunikation oft auf Basis unzureichender oder unpassender Daten fragwürdige Schlussfolgerungen ziehen, während Fragen der Qualität und der Wirkung wenig betrachtet werden. Auch Weitkamp (2015) kritisiert die weit verbreitete Nutzung lediglich deskriptiver Daten bei gleichzeitig wirkungsbezogenen Evaluationsinteressen. King et al. (2015) attestieren, dass es an vielen Stellen an den notwendigen finanziellen und zeitlichen Ressourcen sowie den methodischen Kompetenzen der Praktiker:innen für Wirkungsmessungen und komplexe Evaluationsdesigns fehle. Außerdem wird darauf hingewiesen, dass klassische und häufig eingesetzte Datenerhebungsmethoden wie Papierfragebogen und standardisierte Interviews gerade für die Evaluation dynamischer, interaktiver und weniger formalisierter Wissenschaftskommunikationsformate (wie etwa Festivals) wenig geeignet sind (Grand und Sardo 2017, S. 5).

2 Herausforderungen für die Umsetzung von Evaluationen

Aus einer Gesamtschau der für Deutschland vorliegenden Daten und Eindrücke kristallisieren sich insbesondere drei übergreifende Herausforderungen für die Evaluation von Wissenschaftskommunikation heraus. (1) Es ergeben sich bereits vor Beginn der Evaluationen Probleme durch die fehlende strategische Planung von Wissenschaftskommunikation. (2) Es mangelt bei Evaluationen (insbesondere solchen, die Wirkungsaussagen treffen möchten) oft an passenden

Evaluationsdesigns und geeigneten Datenerhebungsmethoden. (3) Das in der deutschen Wissenschaftskommunikationspraxis generell vorherrschende Bild von Evaluation erschwert einen kollektiven und konstruktiven Lernprozess für die Wissenschaftskommunikation.

2.1 Fehlendes strategisches Vorgehen in der Wissenschaftskommunikation

Klar formulierte Erwartungen an ein Projekt sind essentieller Bestandteil strategischer Projektplanung und bilden die Grundlage für aussagekräftige Evaluationen (Spicer 2017). Idealerweise werden zu Beginn eines Projekts sowohl die Ziele als auch die Zielgruppe des Projekts genau definiert, um anschließend ein möglichst passgenaues Format zu wählen (Besley et al. 2018), das Erfolg bei der Erreichung der Ziele und Zielgruppen verspricht. Evaluationen ermöglichen dann, diese Passgenauigkeit und die Erreichung von Zielen und Zielgruppen zu überprüfen. In der Praxis mangelt es jedoch an Genauigkeit bei jedem dieser Schritte (Phillips et al. 2018; Stilgoe et al. 2014).

Bei der Bestimmung von Zielen bietet die Forschung im Bereich strategischer Kommunikation eine wichtige Orientierung. Dort wird zwischen zwei verschiedenen Arten von Zielen unterschieden. Zum einen spricht die strategische Kommunikation von *goals,* so bezeichnet werden generelle Richtlinien oder abstrakte und übergeordnete Ziele, die nicht notwendigerweise reine Kommunikationsziele sein müssen (Hon 1998, S. 105). Zum anderen ist die Rede von *objectives*, definiert als konkrete Zielsetzungen der Kommunikation, die dazu geeignet sind, die formulierten *goals* zu erreichen (Hallahan 2015, S. 247). In der Praxis wird diese Unterscheidung zwischen *goals* und *objectives* jedoch häufig nicht vorgenommen und es ergeben sich Ungenauigkeiten in der Formulierung der Ziele. Aus der Analyse von Evaluationsberichten im deutschsprachigen Raum ergibt sich der Eindruck, dass Praktiker:innen geübt darin sind, visionsartige *goals* zu formulieren und ihre Projekte in einen größeren Rahmen einzuordnen (Ziegler und Hedder 2020, S. 16 ff.). In Gesprächsrunden mit den Praktiker:innen wurde weiterhin deutlich, dass die Schwierigkeit darin besteht, sich von der Ebene der Visionen wegzubewegen, konkrete *objectives* zu formulieren und diese mit messbaren Indikatoren zu hinterlegen. Symptomatisch dafür ist auch, dass in den analysierten Berichten häufig allgemeine und vage Zielformulierungen wie „ein Bewusstsein zu schaffen“ oder auch „zum Handeln anzuregen“ zu finden sind. Die Erreichung von derart formulierten Zielen lässt sich nur schwer tatsächlich überprüfen und ein solches Vorgehen verleitet dazu, das Augenmerk auf

die Identifizierung irgendeines Effekts zu legen, anstatt die gewünschte Größe dieses Effekts vorab zu definieren und dann zu prüfen (Ziegler und Hedder 2020, S. 19 f.).

Auch bei der Beschreibung der Zielgruppen von Wissenschaftskommunikationsaktivitäten zeigen sich diese Tendenzen zu vagen und ungenauen Formulierungen. In den analysierten Berichten werden meist einfache soziodemographische Merkmale, wie Geschlecht und Alter, zur Zielgruppendefinition herangezogen und nur selten persönlichkeitsbezogene Eigenschaften wie (Vor-)Einstellungen gegenüber Wissenschaft und Forschung (Ziegler und Hedder 2020, S. 19). Zudem gibt beispielsweise nur ein Viertel der Befragten in der Community-Befragung an, den sozioökonomischen Hintergrund zur Zielgruppenbeschreibung zu nutzen (n = 96, Impact Unit 2019, S. 16), obgleich die Forschung zeigt, wie bedeutend dieser für Voreinstellungen zu Wissenschaft und Forschung und damit für das Wirkpotenzial von Wissenschaftskommunikation sein kann (Rutjens et al. 2018). Selbst wenn genauere Zielgruppenbeschreibungen genutzt werden, wird durch Formulierungen wie „hauptsächliche Zielgruppe" (Ziegler und Hedder 2020, S. 19) wiederum der Raum für weitere, nicht genauer definierte Zielgruppen geöffnet. Typisch sind umfassende Bezeichnungen wie *Schüler:innen* oder *die breite Öffentlichkeit*, womit die Chance vergeben wird, eine generische Masse durch Untergruppen zu spezifizieren, die gezielter angesprochen werden könnten (Schäfer und Metag 2021, S. 300).

Darüber hinaus fällt auch die Entscheidung für ein bestimmtes Kommunikationsformat in der Praxis nicht immer basierend auf einer vorherigen Definition von Zielen und Zielgruppen. In der Community-Befragung geben nur 27 % der Befragten an, Formate anhand vorher definierter Ziele auszuwählen, während 73 % ihre Formatwahl darauf zurückführen, „dass jemand eine Idee hat oder ein bestimmtes Format ausprobiert werden soll" (n = 94, Impact Unit 2019, S. 19).

Eine genaue Definition von Zielen und Zielgruppen ist unabdingbar für die informierte Wahl eines Formats, das erfolgsversprechend für die Erreichung der Ziele erscheint. Solch ein strategisches Vorgehen bildet wiederum auch die Grundlage für eine aussagekräftige Evaluation, die überprüfen kann, ob die gewünschte Zielgruppe erreicht wurde und ob das Format tatsächlich geeignet war, die gesteckten Ziele zu erreichen. Doch um solche Einsichten zu ermöglichen, muss auch die Evaluation selbst adäquat geplant werden.

2.2 Defizite bei der Wahl von Evaluationsdesign und -methoden

Bei näherer Betrachtung der analysierten Berichte wird aus den häufig summativen Evaluationsdesigns, den gestellten Evaluationsfragen und den erhobenen Daten deutlich, dass es den Praktiker:innen der Wissenschaftskommunikation oft um die Evaluation von Wirkungen und den Nachweis von Effekten geht (Ziegler und Hedder 2020). Zentral ist hierbei, dass es für die Feststellung einer möglichen Wirkung oder einer erhofften Veränderung zwangsläufig Vergleichswerte braucht. Schließlich kann keine Veränderung – weder eine positive noch eine negative – durch eine Momentaufnahme nachgewiesen werden.

In der Praxis sehen wir jedoch in einigen Fällen eine Diskrepanz zwischen dem Anspruch der Evaluation und der Eignung der gewählten Evaluationsdesigns und -methoden, diesem Anspruch auch gerecht zu werden. Deutlich wird dies, wenn man sich vor Augen führt, dass selbst unter den analysierten Evaluationsberichten ein Viertel der 28 Evaluationen, die nach eigenen Angaben Wirkungen erheben wollen, nur eine einmalige Datenerhebung durchführt (Ziegler und Hedder 2020, S. 24).

Um ein besseres Verständnis von der Wirkung einer Aktivität zu bekommen, wären stattdessen Designs geeignet, die Vorher-Nachher-Vergleiche durchführen oder Kontrollgruppen heranziehen. Während letztere weder in der Community-Befragung noch in den Evaluationsberichten häufig Erwähnung finden, werden in beiden Fällen immerhin von circa einem Drittel Vorher-Nachher-Vergleiche durchgeführt (Ziegler und Hedder 2020, S. 24; Impact Unit 2019, S. 22).

Das in der Praxis übliche Vorgehen, nur einmalig Daten zu erheben, führt gelegentlich dazu, dass die Einschätzung von Veränderungen den Teilnehmenden oder Besucher:innen selbst überlassen wird. So werden diese selbst oder in einigen Fällen auch Dritte (bspw. im Falle von Kindern, deren Eltern oder Lehrkräfte) gefragt, ob sie das Gefühl haben, die Aktivität hätte eine Veränderung ausgelöst. Damit werden die Evaluationsergebnisse von der unrealistischen Annahme abhängig gemacht, dass die Befragten sich an ihr vorheriges Wissen oder ihre vorherigen Einstellungen (im zweiten Fall, den inneren Zuständen anderer Personen) erinnern, diese reflektieren und sinnvoll mit einem späteren Zustand vergleichen können (Jensen und Laurie 2016, S. 158).

Dieses Vorgehen mag auf organisatorischer Ebene ressourcensparend vorkommen, hat allerdings auf inhaltlicher Ebene Folgen für die Aussagekraft der Evaluationsergebnisse. Denn das Einsparen von Erhebungen führt nicht nur zu fehlenden Vergleichspunkten, sondern auch zu blinden Flecken, was die

Perspektiven auf die Aktivitäten angeht: Häufig wird in Evaluationen von Wissenschaftskommunikation nur eine Gruppe von Informationsträger:innen herangezogen – selten mehrere (Ziegler und Hedder 2020, S. 23). Auch befassen sich Evaluationen von Wissenschaftskommunikation in der Regel mit den Perspektiven von Teilnehmenden, meist Bürger:innen, kaum stehen projektinterne Gruppen im Fokus der Evaluation (Ziegler und Hedder 2020, S. 23).

All die genannten Punkte wirken sich negativ auf die Qualität von Evaluationen in der Wissenschaftskommunikation aus. Auch Praktiker:innen sehen diese durchaus kritisch. Nur 6 % von ihnen stimmen laut der Community-Befragung zu, dass Evaluationen in der Wissenschaftskommunikation meistens von guter Qualität sind. Ein Drittel hingegen verneint diese Aussage. Wie eingangs bereits erwähnt, ist mit 39 % der größte Anteil unentschieden und weitere 23 % machen keine Angabe (n = 82, Impact Unit 2019, S. 29). In den Diskussionsrunden mit Praktiker:innen wurde immer wieder deutlich, dass die kurzfristige Planung und Umsetzung von Evaluationen für viele dieser Problematiken ausschlaggebend sein könnte. So kann ein Mangel an Zeit dazu führen, dass bekannte und scheinbar einfach umsetzbare Evaluationsdesigns und Erhebungsmethoden gewählt werden, anstelle komplexerer, aber inhaltlich passenderer. Das zeichnet sich auch in der Community-Befragung ab, in der zwar 93 % die Aussagekraft einer Methode wichtig für die eigene Methodenwahl finden, aber immerhin 87 % angeben, dass eine leichte Plan- und Umsetzbarkeit der Methode für sie ebenfalls hohe Priorität einnimmt. 77 % geben weiterhin an, auch danach auszuwählen, ob eine Methode schnell umsetzbar ist (n = 75, Impact Unit 2019, S. 31).

Auch fehlende Kenntnis und Überforderung können zu Entscheidungen für ungeeignete Methoden beitragen. So stimmen nur 37 % in der Community-Befragung zu, dass sie sich in der Lage sehen, gute Evaluationen für Wissenschaftskommunikationsprojekte zu gestalten (n = 82, Impact Unit 2019, S. 29). Nur 38 % stimmen explizit nicht zu, dass sie es schwierig finden, Interessantes und Relevantes mithilfe von Evaluationen zu erfassen, sodass sich hier wie eingangs erwähnt ein geteiltes Bild ergibt.

Dabei bedarf es keinesfalls in jeder Evaluation komplexer Experimentalstudien oder kostenintensiver Vorher-Nachher-Befragungen. Auch einfach gestaltete deskriptive Evaluationen mit einmaliger Datenerhebung können wichtige Informationen und Einblicke in ein Projekt bieten und beispielsweise für Projekte mit kleineren Budgets, kürzeren Laufzeiten oder experimentellem Charakter angemessener sein als komplexe Erhebungsabfolgen. Wichtig ist aber, dass Evaluationen so geplant werden, dass das Design und die verwendeten Methoden auch tatsächlich dazu geeignet sind, die Evaluationsfragen

zu beantworten. Wenn es nicht möglich ist, ein entsprechendes Evaluationsdesign umzusetzen, gewinnt die Evaluation durch die Anpassung ihrer Fragen mehr Aussagekraft als durch den Versuch, die ursprünglichen Fragen mit dafür ungeeigneten Methoden zu beantworten. Was Praktiker:innen dennoch immer wieder dazu verleiten mag, wirkungsorientierte Fragen zu stellen, könnte sich teilweise durch das vorherrschende Verständnis von Evaluation erklären.

2.3 Verständnis von Evaluation

Bei der Zusammenführung der Ergebnisse aus den für diesen Beitrag herangezogenen Analysen ergibt sich folgender vermeintlicher Widerspruch: Zwar geben in der Community-Befragung 36 % der Befragten an, dass ihre Projekte (fast) immer evaluiert werden und nur 6 % geben an, dass ihre Projekte nie evaluiert werden (n = 96, Impact Unit 2019, S. 21). Gleichzeitig wurden in der Stichwortsuche zum Zweck der Analyse von Evaluationsberichten aus 68 Stichwortkombinationen im Zeitraum von 2000 bis 2019 gerade einmal eine Gesamtzahl von 51 öffentlich zugänglichen Berichten aufgefunden, die 55 Evaluationen vorstellen. Es stellt sich also die Frage, warum scheinbar nur so wenige Evaluationen verfügbar sind und wie mit den anderen Evaluationen nach ihrem Abschluss verfahren wird.

Zwei Interpretationen bieten sich an: Zum einen ist es möglich, dass Praktiker:innen ihre Ergebnisse nicht veröffentlichen, weil sie diese für wenig relevant für Externe halten. Dafür spricht, dass in der Community-Befragung 79 % der Befragten zustimmen, dass ihre Ergebnisse vor allem dazu genutzt werden, im Team die Zusammenarbeit und mögliche Verbesserungen zu reflektieren. Als weitere Verwertungsmöglichkeit darüber hinaus wurde in der Community-Befragung die Weitergabe der Daten zu Forschungszwecken abgefragt – hier stimmen nur 18 % zu (n = 72, Impact Unit 2019, S. 26).

Zum anderen wurde in den Diskussionsrunden mit Praktiker:innen deutlich, dass Evaluationen von vielen als Instrumente zum Erfolgsnachweis oder sogar als Druckmittel verstanden werden. Dazu passt das folgende Muster in den veröffentlichten Berichten: 51 Evaluationen ließen sich mit Blick auf ihre Gestaltung und Funktion im Projekt als summative oder wirkungsorientierte Evaluationen beschreiben, sie zeigen sich also an den Endresultaten des Projekts interessiert.

Lediglich 16 Evaluationen behandeln projektinterne Fragen und zeigen sich damit an den Projektprozessen interessiert[5] (Ziegler und Hedder 2020, S. 21).

Evaluationen wurden in den Diskussionsrunden darüber hinaus immer wieder als langer Arm von Vorgesetzten oder Förderinstitutionen interpretiert, die zukünftige Finanzierung an Erfolge knüpfen. Auch für diese Vermutung finden sich einige Hinweise in der Community-Befragung. Die Hälfte der Befragten stimmt zu, dass Evaluationen manchmal dazu genutzt werden, um Erfolge zu belegen, obwohl dafür nicht die richtigen Daten vorliegen. 49 % stimmen zu, dass Evaluationen vor allem dazu dienen, die eigene Arbeit vor Anderen zu belegen. Demgegenüber stimmen nur 39 % zu, dass Evaluationsergebnisse in die Neu- und Weiterentwicklung von Projekten einfließen (n = 82, Impact Unit 2019, S. 30). So wäre es also durchaus möglich, dass Ergebnisse bewusst unter Verschluss gehalten werden, um Probleme oder ausbleibende Erfolge nicht öffentlich und das eigene Projekt damit angreifbar zu machen (Nothhaft und Stensson 2019).

Ein derartiges Verständnis von Evaluation scheint die Arbeitsrealität für Praktiker:innen der Wissenschaftskommunikation in Deutschland nachhaltig zu bestimmen. Evaluationen werden dabei aus verschiedenen Gründen angestoßen und ihre Ergebnisse sollen legitimerweise verschiedenen Zwecken dienen (siehe auch Niemann et al. in diesem Band; Volk in diesem Band) – was auch bedeuten kann, dass ihre Ergebnisse nicht in allen Fällen für die Öffentlichkeit bestimmt sind oder bestimmt sein können. Ergänzend ist auch anzuführen, dass beispielsweise externe Dienstleistende, die Evaluationen durchführen, häufig nicht die Entscheidung über die Ergebnisveröffentlichung treffen. Vielmehr endet deren Arbeit oftmals mit der Übergabe der Ergebnisse an die auftraggebende Person oder Einrichtung, welche die weitere Nutzung der Ergebnisse bestimmt.

Insgesamt ist eine Folge dieses Verständnisses aber, dass wichtige Erkenntnisse aus Evaluationen der weiteren Community von Praktiker:innen vorenthalten werden. So können Projekte nicht aus den Fehlern oder von den Erfolgsfaktoren Anderer lernen und Evaluation wird zu einem Schreckgespenst anstatt zu einem kollektiven und konstruktiven Lernprozess. Um Evaluation zu einem solchen Lernprozess zu machen, muss ein Umdenken stattfinden. Hoffentlich kann unter anderem dieser Band einen Beitrag dazu leisten und zu einer anderen Evaluationspraxis der Zukunft beitragen.

---

[5] Evaluationen mit mehreren Erkenntnisinteressen und Evaluationsfragen können natürlich sowohl prozessorientierte als auch wirkungsorientierte Anteile mitbringen. Aus diesem Grund übersteigt die Zahl der prozessorientierten und wirkungsorientierten Evaluationen die Gesamtzahl der analysierten Evaluationen.

3 Ausblick

In der zukünftigen Praxis der Wissenschaftskommunikation sollten Evaluationen als konstruktive, kollektive Lernprozesse verstanden werden, von denen Wissenschaftskommunikation insgesamt, und damit letztendlich auch die Gesellschaft, profitieren kann. Wissenschaftskommunikation ist divers – in ihren Formaten, Kanälen und auch Akteur:innen. Das Ziel von Evaluationen kann es daher nicht sein, *one-size-fits-all*-Indikatoren zu entwickeln, die ohne Berücksichtigung der konkreten Umstände der Kommunikation zur unreflektierten Erfolgsmessung und -dokumentation verwendet werden. Stattdessen müssen Evaluationen dieser Verschiedenheit Rechnung tragen und durch diverse Methoden und einen kritischen Reflexionsprozess Einblicke und Schlussfolgerungen für eine Bandbreite an Ansätzen zulassen. Hier kann dieser Band einen wichtigen Anstoß und Einblicke in die diversen Möglichkeiten zu deren Umsetzung geben. Inspiration geben auch Bestrebungen zur Professionalisierung von Evaluationen in anderen Bereichen[6], die sich auf die zukünftige Entwicklung der Wissenschaftskommunikation übertragen lassen.

So verstanden und praktiziert können Evaluationen auch über den einzelnen Lernprozess für konkrete Aktivitäten und Projekte hinaus einen entscheidenden Beitrag leisten. Evaluationen können unter anderem eine wichtige Basis für Diskussionen über die Qualität von Wissenschaftskommunikation bilden. Zum einen erheben sie wichtige Primärdaten. Zum anderen erlauben sie eine Prüfung genereller Ideen und Vorstellungen unter realen Bedingungen. So können strategische Ziele der Wissenschaftskommunikation stärker mit der Wissenschaftskommunikationspraxis abgestimmt werden, was realistische, für die Praxis erreichbare Zielformulierungen erleichtert. Wenn dieser Raum der realistischen Möglichkeiten abgesteckt ist, kann auch eine konstruktive und normative Diskussion über Wissenschaftskommunikation geführt werden. Größere Klarheit darüber, was Wissenschaftskommunikation wirklich erreichen kann, stellt die Diskussion darüber, was sie erreichen soll, auf eine solide Basis.

Aussagekräftige Evaluationen sind nicht nur für die Wissenschaftskommunikationspraxis wichtig, sondern leisten auch einen Beitrag dazu, dass Fördergelder sinnvoll und zielgerichtet eingesetzt werden. Dabei geht es nicht um die bloße Messung von Erfolgen, sondern ganz entscheidend auch darum, dass

[6] Beispielhaft zu nennen ist hier der Arbeitskreis Professionalisierung der DeGEval – Gesellschaft für Evaluation e. V.

Ziele von fördernden Institutionen klar und realistisch formuliert werden, sodass Projekte danach ausgerichtet werden können.

Daraus ergeben sich wiederum umfassende Möglichkeiten für die Praxis. Eine strategische Planung von Wissenschaftskommunikation, wie wir sie in diesem Beitrag entworfen haben, ist informierter möglich, wenn Praktiker:innen eine Grundlage zur Einschätzung haben, welche Ziele und Zielgruppen in der Wissenschaftskommunikation (leichter oder schwieriger) erreichbar und welche Formate in der Praxis tatsächlich geeignet sind. Damit nicht jedes Projekt diesen Lernprozess aufs Neue durchlaufen muss, ist umso entscheidender, dass Evaluationen als kollektive Reflexionsprozesse der gesamten Community verstanden werden. So erlauben es Evaluationen, Projekte weiterzuentwickeln und zu verbessern, und sind neben der Forschung zu Wissenschaftskommunikation eine weitere wichtige Informationsquelle, um Wissenschaftskommunikation zu verstehen und zu verbessern.

Gleichzeitig dürfen Evaluationen nicht dazu verleiten, Wissenschaftskommunikation immer mehr auf wenige bewährte Formate zu verengen. Ganz im Gegenteil, Evaluationen sollen den Raum für Kreativität und Experimentierfreude in der Wissenschaftskommunikation öffnen. Wenn neue Formate systematisch getestet und durch sorgfältige Evaluationen begleitet werden, können sie den Kanon geeigneter Formate stetig erweitern und das Instrumentarium der Wissenschaftskommunikation konstruktiv ergänzen.

Literatur

Besley JC, Dudo A, Yuan S (2018) Scientists' views about communication objectives. Public Unders Sci 27(6):708–730. https://doi.org/10.1177/0963662517728478

Besley JC, O'Hara K, Dudo A (2019) Strategic science communication as planned behavior: understanding scientists' willingness to choose specific tactic. PLoS ONE 14(10):e0224039. https://doi.org/10.1371/journal.pone.0224039

Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (2017) Das Forschungsfeld Wissenschaftskommunikation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer M (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-12898-2_1

Bundesministerium für Bildung und Forschung (BMBF) (2019) Grundsatzpapier des Bundesministeriums für Bildung und Forschung zur Wissenschaftskommunikation. https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/1/24784_Grundsatzpapier_zur_Wissenschaftskommunikation.pdf. Zugegriffen: 17. November 2022

Dudo A, Besley JC (2016) Scientists' prioritization of communication objectives for public engagement. PLoS ONE 11(2):e0148867. https://doi.org/10.1371/journal.pone.0148867

Fischhoff B (2019) Evaluating science communication. Proc Natl Acad Sci 116(16):7670–7675. https://doi.org/10.1073/pnas.1805863115
Grand A, Sardo AM (2017) What works in the field? Evaluating informal science events. Front Commun 2(22). https://doi.org/10.3389/fcomm.2017.00022
Hallahan K (2015) Organizational goals and communication objectives in strategic communication. In: Holtzhausen D, Zerfass A (Hrsg) The Routledge Handbook of Strategic Communication, 1. Aufl. Routledge, London, S 244–266
Hon LC (1998) Demonstrating effectiveness in public relations: goals, objectives, and evaluation. J Public Relat Res 10(2):103–135. https://doi.org/10.1207/s1532754xjprr1002_02
Impact Unit (2019) Evaluation and impact in science communication: results of a community survey 2019. Wissenschaft im Dialog, Berlin. https://www.wissenschaft-im-dialog.de/fileadmin/user_upload/Projekte/Impact_Unit/Dokumente/Impact_Unit_2019_Community_Survey.pdf . Zugegriffen: 26. Dezember 2021
Jensen EA (2014) The problems with science communication evaluation. JCOM, 13(1), C04. https://doi.org/10.22323/2.13010304
Jensen EA (2020) Why impact evaluation matters in science communication: or, advancing the science of science communication. In: Weingart P, Joubert M, Bankole F (Hrsg) Science communication in South Africa. African Minds, Kapstadt. https://doi.org/10.5281/zenodo.3557213
Jensen EA, Gerber A (2020) Evidence-Based science communication. Front Commun 4(78) https://doi.org/10.3389/fcomm.2019.00078
Jensen EA, Laurie C (2016) Doing real research: a practical guide to social research. SAGE, London
King H, Steiner K, Hobson M, Robinson A, Clipson H (2015) Highlighting the value of evidence-based evaluation: pushing back on demands for 'impact'. JCOM 14(2):A02. https://doi.org/10.22323/2.14020202
Mannino I, Bell L, Costa E, Di Rosa M, Fornetti A, Franks S, Iasillo C, Maiden N, Olesk A, Pasotti J, Renser B, Roche J, Schofield B, Villa R, Zollo F (2021) Supporting quality in science communication: insights from the QUEST project. JCOM 20(03):A07. https://doi.org/10.22323/2.20030207
National Academies of Sciences, Engineering, and Medicine (2017). Communicating science effectively: a research agenda. The National Academies Press, Washington, DC. https://www.nap.edu/read/23674/chapter/1. Zugegriffen: 17. November 2022
Nothhaft H, Stensson H (2019) Explaining the measurement and evaluation stasis: a thought experiment and a note on functional stupidity. JCOM 23(03):213–227. https://doi.org/10.1108/JCOM-12-2018-0135
Olesk A, Renser B, Bell L et al (2021) Quality indicators for science communication: results from a collaborative concept mapping exercise. JCOM 20(03):A06. https://doi.org/10.22323/2.20030206
Pellegrini G (2021) Evaluating science communication. In: Bucchi M, Trench B (Hrsg) Routledge handbook of public communication of science and technology. Routledge, Abington
Peterman K, Verbeke M, Nielsen K (2020) Looking back to think ahead: reflections on science festival evaluation and research. Visitor Stud 23(2):205–217. https://doi.org/10.1080/10645578.2020.1773709

Phillips T, Porticella N, Constas M, Bonney R (2018) A framework for articulating and measuring individual learning outcomes from participation in citizen science. Citizen Sci Theory Pract 3(2):1–19. https://doi.org/10.5334/cstp.126

Rutjens BT, Heine SJ, Sutton RM, Harreveld F (2018) Chapter three – attitudes towards science. Adv Exp Soc Psychol 57:125–165. https://doi.org/10.1016/bs.aesp.2017.08.001

Schäfer MS, Metag J (2021) Audiences of science communication between pluralization, fragmentation and polarization. In: Bucchi M, Trench B (Hrsg) Handbook of public communication of science and technology. Routledge, London, S 291–304

Schäfer MS, Kristiansen S, Bonfadelli H (2015) Wissenschaftskommunikation im Wandel: Relevanz, Entwicklung und Herausforderungen des Forschungsfeldes. In: Bonfadelli H, Schäfer MS, Kristiansen S (Hrsg) Wissenschaftskommunikation im Wandel. Herbert von Halem Verlag, Köln, S 10–42

Scheufele DA, Krause NM, Freiling I, Brossard D (2021) What we know about effective public engagement on CRISPR and beyond. Proc Natl Acad Sci USA 118(22):e2004835117. https://doi.org/10.1073/pnas.2004835117

Spicer S (2017) The nuts and bolts of evaluating science communication activities. Semin Cell Dev Biol 70:17–25. https://doi.org/10.1016/j.semcdb.2017.08.026

Stilgoe J, Lock SJ, Wilsdon J (2014) Why should we promote public engagement with science? Public Underst Sci 23(1):4–15. https://doi.org/10.1177/0963662513518154

Weitkamp E (2015) Between ambition and evidence. JCOM 14(2):E. https://doi.org/10.22323/2.14020501

Wicke N (2021) Eine Frage der Erwartungen? Publizistik 67:51–84. https://doi.org/10.1007/s11616-021-00701-z

Ziegler R, Hedder IR (2020) Evaluationspraktiken der Wissenschaftskommunikation. Eine Betrachtung veröffentlichter Evaluationsberichte im Deutschsprachigen Raum. Wissenschaft im Dialog. https://www.wissenschaft-im-dialog.de/fileadmin/user_upload/Projekte/Impact_Unit/Dokumente/210701_Ergebnisbericht_Evaluationspraktiken.pdf. Zugegriffen: 3. Januar 2022

Ziegler R, Hedder IR, Fischer L (2021) Evaluation of science communication: current practices, challenges, and future implications. Front Commun 6(669744). https://doi.org/10.3389/fcomm.2021.669744

Ricarda Ziegler ist Leiterin des Bereichs Qualität & Transfer bei Wissenschaft im Dialog (WiD) – der deutschen Organisation für Wissenschaftskommunikation. Sie verantwortet dort u. a. die Impact Unit, die sich Fragen der Wirkung und Evaluation von Wissenschaftskommunikation widmet. Außerdem leitet sie das bevölkerungsrepräsentative Wissenschaftssurvey Wissenschaftsbarometer. Ricarda Ziegler hat einen Hintergrund in der Politikwissenschaft.

Imke Hedder arbeitet für die deutsche Organisation für Wissenschaftskommunikation Wissenschaft im Dialog (WiD) im Bereich Qualität & Transfer. Als Teil der Impact Unit führte sie Analysen zur Evaluationspraxis in der deutschen Wissenschaftskommunikation durch und entwickelte Evaluationstools für Praktiker:innen. Inzwischen ist sie für die Evaluation verschiedener WiD-Projekte zuständig.

Liliann Fischer ist stellvertretende Projektleiterin und wissenschaftliche Mitarbeiterin der Transfer Unit bei Wissenschaft im Dialog (WiD). In diesem Kooperationsprojekt mit der Berlin-Brandenburgischen Akademie der Wissenschaften widmet sie sich der Beförderung eines konstruktiven Austauschs von Wissenschaftskommunikationspraxis und -forschung. Neben ihrer Tätigkeit bei Wissenschaft im Dialog promoviert sie an der Universität Passau zu Selbstverständnissen in der Wissenschaftskommunikation.



Evaluation der Wissenschaftskommunikation: Modelle, Stufen, Methoden

Sophia C. Volk

Zusammenfassung

Wissenschaftskommunikation kann unterschiedliche Ziele verfolgen, sich an unterschiedliche Zielgruppen richten und dabei unterschiedliche Maßnahmen einsetzen. Ob sie die beabsichtigten Wirkungen erzielt, lässt sich erst durch eine systematische Evaluation feststellen. Dieser Beitrag führt in die Logik von Evaluation ein, indem es grundlegende Evaluationsmodelle, -stufen und -objekte erklärt. Es stellt ein integriertes Evaluationsmodell für die Wissenschaftskommunikation vor, das a) Evaluation als letzte Phase in einem übergeordneten Zyklus begreift (Situationsanalyse, Planung, Umsetzung), b) zwei Evaluationsformen (summativ, formativ) dazu in Bezug setzt, c) vier Evaluationsstufen (Inputs, Outputs, Outcomes, Impacts) unterscheidet und d) verschiedene Evaluationsobjekte (Projekt, Kampagne, Programme) und Zeithorizonte (kurz-, mittel-, langfristig) berücksichtigt. Der Beitrag gibt zudem einen Überblick über sozialwissenschaftliche und betriebswirtschaftliche Methoden und typische Kennzahlen für die Evaluation der Wissenschaftskommunikation.

S. C. Volk (✉)
Institut für Kommunikationswissenschaft und Medienforschung, Universität Zürich, Zürich, Schweiz
E-Mail: s.volk@ikmz.uzh.ch

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_3

1 Einleitung

Wissenschaftskommunikation gewinnt zunehmend als wissenschaftlicher Forschungszweig und als Berufsfeld an Bedeutung. Dies zeigt sich beispielsweise in einer Zunahme von Ressourcen für Wissenschaftskommunikation oder in der Ausdifferenzierung von verschiedenen Maßnahmen und Formaten der Wissenschaftskommunikation: Von Science Slams über Podcasts zu Schülerlaboren, Science Cafés, Hackathons oder Langen Nächten der Wissenschaft (Ziegler et al. 2021). Mit solchen Maßnahmen können viele unterschiedliche Ziele verfolgt werden, etwa das Wissen zu bestimmten Themen bei wissenschaftsfernen Zielgruppen zu steigern, die Akzeptanz oder das Vertrauen in neue Technologien zu stärken oder aber ein positives Image von wissenschaftlichen Institutionen aufzubauen (Weingart und Joubert 2019). Ob diese Ziele mithilfe von Wissenschaftskommunikation erreicht werden konnten und ob vorhandene Ressourcen und konkrete Kommunikationsmaßnahmen dabei effektiv und effizient eingesetzt wurden, lässt sich erst durch eine systematische Evaluation beantworten. Die Bedeutung von Evaluation wird erst seit kürzerer Zeit in der Forschung und Praxis der Wissenschaftskommunikation diskutiert (Jensen 2014). Dieser Beitrag führt in die Logik von Evaluation ein, indem es grundlegende Evaluationsmodelle, -stufen und -objekte erklärt und daraufhin ein integriertes Evaluationsmodell für die Wissenschaftskommunikation vorstellt. Es gibt darüber hinaus einen Überblick über sozialwissenschaftliche und betriebswirtschaftliche Methoden und typische Kennzahlen für die Evaluation der Wissenschaftskommunikation, die idealerweise miteinander kombiniert werden. Schließlich werden aktuelle Herausforderungen für Wissenschaft und Praxis skizziert.

2 Logik von Evaluationen

Theorien, Modelle und Methoden für die Evaluation werden seit Jahrzehnten in verschiedenen Disziplinen erforscht – etwa in der öffentlichen Verwaltungswissenschaft, der Bildungsforschung, der internationalen Entwicklungsforschung oder Managementforschung – und in verschiedenen Berufsfeldern angewandt (Macnamara 2018). Die Kommunikationswissenschaft hat sich mit Evaluationsforschung nur am Rande beschäftigt; erst seit den 1990er Jahren lässt sich eine zunehmende Auseinandersetzung mit Fragestellungen und Praktiken der Evaluation im Bereich der Strategischen Kommunikationsforschung feststellen (Volk 2016). Das ist für die Wissenschaftskommunikationsforschung und -praxis von Vorteil, denn Schlüsselkonzepte und Theorien aus der Evaluationsforschung

wurden bereits auf das Forschungsfeld (Strategische) Kommunikation übertragen, etwa die Program Theory (z. B. Frechtling 2015; Rossi et al. 2004), das Prinzip von Logikmodellen (Knowlton und Phillips 2013) und die Theory of Change (z. B. Clark und Taplin 2012; Funnell und Rogers 2011). Da die Wissenschaftskommunikation neben dialogischen Zielen auch strategische oder persuasive Ziele verfolgt (Besley und Dudo, 2022) – wenn auch andere als etwa Unternehmen, Parteien oder NGOs – lassen sich die grundlegende Logik von Evaluation, die Stufen von Evaluation und die Objekte von Evaluation mit einigen Anpassungen adaptieren (Raupp und Osterheider 2019).

2.1 Formen von Evaluation

Unter Evaluation (lateinisch von „valere“: wert sein) wird allgemein die Bewertung und Überprüfung von Kommunikationsaktivitäten im Hinblick auf gesetzte Ziele verstanden. Was die Evaluationsforschung von der Rezeptions- und Wirkungsforschung der Wissenschaftskommunikation unterscheidet, ist die Tatsache, dass nicht allgemeine Effekte von Kommunikation untersucht werden, sondern Wirkungen im Hinblick auf vorab definierte Kommunikationsziele (Pellegrini 2021; Raupp 2017). Evaluation stellt dabei die letzte Phase in einem Zyklus dar, der aus vier Kernelementen besteht: *Situationsanalyse* (Bedarfsermittlung), *Planung* (Strategiefindung), *Umsetzung* (Strategieimplementierung) und *Evaluation* (Zielerreichung). Im Rahmen der Planungsphase müssen die durch Wissenschaftskommunikation zu erreichenden *Ziele* (z. B. Aufmerksamkeit, Vermittlung von Wissen, Verständigung, Partizipation) eindeutig festgelegt werden, etwa entlang der SMART-Formel (Akronym für: spezifisch, messbar, attraktiv, realistisch und terminlich fixiert) (Volk und Zerfaß 2022). Neben Zielen sollten auch die *Zielgruppen* konkretisiert werden, z. B. anhand ihrer demografischen Merkmalen, Einstellungen gegenüber bzw. Interesse an Wissenschaft oder ihrem wissenschaftsbezogenen Mediennutzungsverhalten (Ziegler et al. 2021). Ein Beispiel könnte lauten: „Im dritten Quartal soll durch die Kinder-Uni die Bekanntheit der Universität X in der wissenschaftsfernen Bevölkerung Y im Raum Z um 5 Prozentpunkte gesteigert werden.“ Ob das Ziel durch den Einsatz von Wissenschaftskommunikation verwirklicht werden konnte, lässt sich im Rahmen der Evaluation durch einen Vergleich der angestrebten Ziele (Soll-Werte) und realisierten Ziele (Ist-Werte) bestimmen.

In der Literatur werden meist zwei Formen von Evaluationen unterschieden: „Formative“ und „summative“ Evaluation (vgl. van den Bogaert in diesem Band). Die *summative* Evaluation untersucht *nach* der Umsetzung bzw. Strategieimplementierung, inwieweit die Kommunikationsaktivitäten ihre Ziele erreicht

haben; sie dient zur Rechenschaftslegung und ermöglicht Lernprozesse, um künftige Planungen zu verbessern. Die *formative* Evaluation hingegen findet *vor* und *begleitend* zur Umsetzung von Kommunikationsmaßnahmen statt. Im Vorfeld dient sie der Analyse der Ausgangssituation und der Generierung von Insights für die Planung, etwa durch die Analyse von Interessen und Kanalpräferenzen des Publikums. Während der Umsetzung dient sie der kontinuierlichen Beobachtung (im Sinne von Monitoring) und Optimierung bzw. Anpassung von Kommunikationsprozessen; manche Autor:innen bezeichnen diese prozessbegleitende Form von Evaluation daher auch als *prozessuale* Evaluation (z. B. Buhmann und Likely 2018; Pellegrini 2021; Rice und Atkin 2013; Watson und Noble 2014).

2.2 Evaluationsmodelle, -stufen und -objekte

Die Entwicklung von Modellen für die Evaluation reicht viele Jahrzehnte zurück. Für die Evaluation von Strategischer Kommunikation liegen verschiedene, wissenschaftlich fundierte und praxiserprobte Evaluationsmodelle vor (vgl. Macnamara und Gregory 2018). Was unterschiedliche Evaluationsmodelle gemein haben ist, dass sie trotz unterschiedlichen Terminologien mehr oder weniger der Struktur von Logikmodellen ähneln. Bei Logikmodellen handelt es sich um vereinfachte Darstellungen verschiedener Stufen von Wirkungen. Um vermutete (kausale) Beziehungen zwischen den einzelnen Stufen zu erklären, wird häufig auf Annahmen aus der Program Theory und der Theory of Change zurückgegriffen (z. B. Clark und Taplin 2012; Frechtling 2015). In ihrer grundlegendsten Form unterscheiden diese Logikmodelle zwischen folgenden vier Stufen: *Inputs* (die Ressourcen, die in eine Aktivität fließen), *Outputs* (die Produkte bzw. Aktivitäten, die daraus resultieren), *Outcomes* (die kurz- und mittelfristigen Veränderungen, die aus den Aktivitäten resultieren) und *Impacts* (die in der Regel langfristigen Ergebnisse z. B. auf gesellschaftlicher, erzieherischer, ökologischer, demokratischer Ebene).

Ein Versuch, verschiedene Evaluationsmodelle im Bereich der Strategischen Kommunikation zu standardisieren und zu vereinheitlichen, wurde in den letzten Jahren von der International Association for the Measurement and Evaluation of Communication in Form des *Integrated Evaluation Framework* (IEF) vorgenommen (AMEC 2016), das mittlerweile in 20 Sprachen übersetzt wurde. Im deutschsprachigen Diskurs hat das sogenannte *DPRG/ICV-Wirkungsstufenmodell* (DPRG und ICV 2011) eine breitere Resonanz erfahren. Die Logik des DPRG/ICV-Wirkungsstufenmodell wurde bereits von Raupp und Osterheider (2019) oder Scheuerle et al. (2017) und Scheuerle (2020) für die

Wissenschaftskommunikation adaptiert und findet sich in ähnlicher Form auch in der englischsprachigen Literatur (Pellegrini 2021). Demnach lässt sich Wissenschaftskommunikation entlang von vier Wirkungsstufen evaluieren:

1. *Inputs* umfassen die Ressourcen, die für die Vorbereitung und Durchführung von Projekten und Maßnahmen der Wissenschaftskommunikation benötigt werden (z. B. Zeit, finanzielle und personelle Ressourcen); die Input-Phase stellt die Brücke zwischen Planung und Durchführung dar.
2. *Outputs* umfassen die Leistungen, die durch Maßnahmen der Wissenschaftskommunikation, geschaffen und von Zielgruppen genutzt werden. Es können *interne* Outputs (z. B. Anzahl von Pressemitteilungen, Social Media Posts oder Events) und *externe* Outputs (z. B. Präsenz in der Medienberichterstattung, Reichweite auf Social Media, Visits auf Webseite) unterschieden werden.
3. *Outcomes* umfassen kognitive, affektive und konative sowie physiologische Wirkungen, die durch Maßnahmen der Wissenschaftskommunikation bei Zielgruppen geschaffen werden. Sie können in *direkte* Outcomes (z. B. Aufmerksamkeit, Recall, Recognition) und *indirekte* Outcomes (z. B. Interesse, Wissen, Lernen, Einstellungsveränderungen, Emotionen, Verhaltensänderungen) unterschieden werden; hierbei handelt es sich um kurz- und mittelfristige Wirkungen.
4. *Impacts* umfassen den langfristigen Wert, der auf gesellschaftlicher Ebene (z. B. Vertrauen in die Wissenschaft, Akzeptanz von neuen Technologien) oder institutioneller Ebene (z. B. Image, öffentliche Finanzierung, Legitimation, Vertrauen in die Institution) geschaffen wird. Hierbei handelt es sich um langfristige Wirkungen, die mithilfe von mehreren Messzeitpunkten erhoben werden können.

Die skizzierte Evaluationslogik kann dabei auf unterschiedlichen Ebenen der Wissenschaftskommunikation zum Einsatz kommen: Als Evaluationsobjekte können einzelne Projekte bzw. Kommunikationsprodukte (z. B. TikTok Video, Pressekonferenz; Fokus: eher kurzfristig), größere Kommunikationskampagnen (z. B. cross-mediale Kampagne zu Citizen Science; Fokus: eher kurz- und mittelfristig) oder umfangreiche Kommunikationsprogramme (z. B. wiederkehrende Kinder-Uni, Science Festival; Fokus: eher langfristig) im Hinblick auf ihre Zielerreichung bewertet werden. Diese sind ineinander „verschachtelt", d. h. einzelne Projekte sind oft Elemente von größeren Kampagnen, die wiederum Teil größerer Programme sind (Buhmann und Volk 2022). Über die Kommunikation hinaus können auch Kommunikationsteams (z. B. mit Verantwortlichkeit für interne Wissenschaftskommunikation) oder ganze Kommunikationsabteilungen in

Wissenschaftsorganisationen (z. B. Museum, Universität) entlang der vier Stufen evaluiert werden. Der Grad an Komplexität und der Zeithorizont (kurz-, mittel- und langfristig) der Evaluation nimmt mit jeder Einheit und jeder Stufe zu.

Führt man die obigen Überlegungen zusammen, lässt sich für die Evaluation der Wissenschaftskommunikation ein integriertes Evaluationsmodell (in Anlehnung an Buhmann und Likely 2018; Buhmann und Volk 2022) vorschlagen, das a) Evaluation als letzte *Phase* in einem übergeordneten Zyklus begreift (Situationsanalyse, Planung, Umsetzung), b) zwei *Evaluationsformen* (summativ, formativ/prozessual) dazu in Bezug setzt, c) vier *Evaluationsstufen* (Inputs–Outputs–Outcomes–Impacts) in der Phase der Umsetzung differenziert und (d) verschiedene *Evaluationsobjekte* (Projekt, Kampagne, Programme) und Zeithorizonte (kurz-/mittel-/langfristig) berücksichtigt (Abb. 1).

Evaluationsmodelle werden häufig implizit oder explizit mit der Annahme hinterlegt, dass Outputs, Outcomes und Impacts in einer *Ursache-Wirkungs-Beziehung* stünden, es also kausale Wirkungszusammenhänge zwischen den Aktivitäten der Wissenschaftskommunikation und den Wirkungsstufen gäbe (Volk und Zerfaß 2022). Während es in der Praxis durchaus pragmatisch zielführend sein kann, *logisch anzunehmende* Wirkungsketten zu vereinbaren, ist es aus einer wissenschaftlichen Sicht problematisch, wenn gemessene Wirkungen kausal auf einzelne Maßnahmen der Wissenschaftskommunikation zurückgeführt werden (vgl. Raupp und Osterheider 2019): Denn zum einen ist Wissenschaftskommunikation häufig nur ein Faktor unter vielen anderen Einflussfaktoren (z. B. Eintrittspreis, Entfernung zu einer Ausstellung), der Einstellungen oder

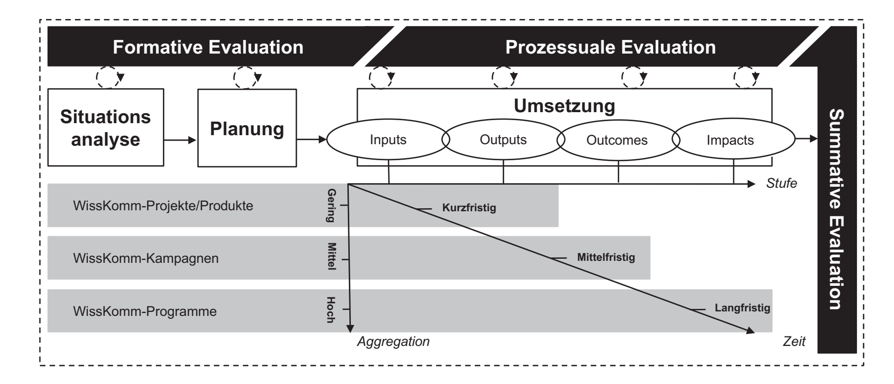


Abb. 1 Evaluationsmodell für die Wissenschaftskommunikation

Verhaltensabsichten beeinflussen kann; alternative Einflussfaktoren werden empirisch aber meist nicht gemessen und Kommunikationswirkungen treten oft zeitversetzt auf. Daher lassen sich Änderungen nicht eindeutig auf Wissenschaftskommunikation zurückführen bzw. *verursachungsgerecht* zurechnen. Zum anderen werden in der Praxis aus forschungsökonomischen Gründen meist keine experimentellen Testverfahren mit Kontrollgruppen oder Vorher-Nachher-Designs bzw. Langzeiterhebungen genutzt, sodass keine empirischen Aussagen über Kausalitäten getroffen werden können (Jensen 2015; King et al. 2015; Ziegler et al. 2021). Daher ist das Wissen um die Grenzen und Limitationen von Evaluationen für die Analyse und Interpretation von Daten essenziell.

3 Evaluationsmethoden für die Wissenschaftskommunikation

Im Zuge der Evaluation kann ein breites Spektrum an Methoden für die Bewertung von einzelnen Projekten, Kampagnen oder Programmen der Wissenschaftskommunikation zum Einsatz kommen. Die Ergebnisse von Messungen werden als Kennzahlen bezeichnet, die als (quantitative oder qualitative) Leistungsindikatoren einen Soll-Ist-Vergleich zwischen den angestrebten Zielen und den tatsächlichen Ergebnissen ermöglichen. Kennzahlen, die kritische und strategisch relevante Informationen zusammenfassen, werden auch als „Key Performance Indicators“ (KPIs) bezeichnet (Volk und Zerfaß 2022).

Unterschiedliche Methoden und Kennzahlen lassen sich entlang der Evaluationsstufen systematisieren: Auf der *Output- und Outcome-Ebene* stehen die Inhalte von Kommunikationsangeboten, Nutzungsmuster und Kommunikationswirkungen bei den Zielgruppen bzw. in der Gesellschaft im Vordergrund. Auf der vorgelagerten *Input-Ebene* und der nachgelagerten *Impact-Ebene* hingegen stehen die Effizienz beim Ressourceneinsatz und die geschaffenen immateriellen und ggf. materiellen Werte im Fokus. Dementsprechend sind neben den klassischen sozialwissenschaftlichen Forschungsmethoden auch betriebswirtschaftliche Methoden für eine systematische Evaluation notwendig (Raupp und Osterheider 2019). Eine ganzheitliche Betrachtung von Inputs bis zu Impacts ist sinnvoll, da Kennzahlen auf der Outcome-Ebene zu Wirkungen bei Zielgruppen wenig darüber aussagen, ob die gewählten Formate der Wissenschaftskommunikation auch effizient und ressourcenschonend umgesetzt wurden oder ein langfristiger Beitrag für die gesamte Wissenschaftsorganisation (z. B. Museum, Universität) oder einzelne Akteur:innen (z. B. Wissenschaftler:innen) geschaffen wurde. Aussagen über

Effizienz und Impact von spezifischen Maßnahmen oder Projekten in der Wissenschaftskommunikation setzen natürlich grundlegend voraus, dass Ziele im Rahmen der strategischen Planung klar definiert und Wirkungen auf den vorgelagerten Output- und Outcome-Stufen gemessen werden. Genau dies stellt jedoch in der Praxis der Wissenschaftskommunikation derzeit eine Herausforderung dar (Ziegler et al. 2021).

3.1 Sozialwissenschaftliche Methoden

Für die Messung von Kommunikationswirkungen auf der *Output- und Outcome-Ebene* bietet sich die Nutzung der gesamten Bandbreite der quantitativen und qualitativen sozialwissenschaftlichen Forschungsmethoden an. Darunter fallen alle in diesem Buch vorgestellten Methoden: Befragungen, Beobachtungen, Nutzungs-Datenanalysen, Inhaltsanalysen, physiologische Messungen, Testverfahren oder experimentelle Herangehensweisen (vgl. Grundlagenbeiträge in diesem Band). Je nach Erkenntnisinteresse lassen sich verschiedene Methoden im Sinne von Mixed Methods miteinander kombinieren (vgl. Gabriel, Kessel, Quast und Roth in diesem Band).

Auf der *internen Output-Ebene* können zunächst einfache Auszählungen genutzt werden, um etwa die in einem bestimmten Zeitraum selbst erstellte Anzahl von Social-Media-Posts, Pressemitteilungen, Events usw. auszuwerten. Auf der *externen Output-Ebene* können mithilfe von Nutzungsdatenanalysen (z. B. Webtracking) Aussagen über die Anzahl von Visitors, Page Impressions oder die durchschnittliche Verweildauer von User:innen auf der eigenen Webseite getroffen werden (vgl. Bruckermann und Greving in diesem Band). Für Social-Media-Kanäle kann die Reichweite von Posts, Tweets oder Stories ermittelt werden. Diese Kennzahlen erlauben Rückschlüsse auf die Anzahl der potenziellen Kontakte mit der Zielgruppe, sagen aber nichts darüber aus, ob die Zielgruppe die Inhalte tatsächlich gesehen oder gehört hat oder was sie daraufhin gedacht und getan hat. Ferner lassen sich Inhaltsanalysen (z. B. Medienresonanzanalysen, Clippings) nutzen, um die Präsenz der eigenen Formate in der Medienberichterstattung zu untersuchen und Aussagen über die Tonalität und den Share of Voice treffen zu können (vgl. Kessler und Wicke in diesem Band; siehe auch Raupp und Vogelgesang 2009). Analog dazu können Kommentare oder Posts auf Social-Media-Kanälen inhaltsanalytisch, etwa in Bezug auf Themen (z. B. Hashtags) und Sentiments (z. B. positiv, negativ, neutral), untersucht werden. Auf der *direkten Outcome-Ebene* lassen sich mithilfe von Social-Media-Analytics die Anzahl an Likes, Shares, Followers, Retweets etc. verfolgen.

Diese liefern nützliche Indikatoren für ein Potenzial für kommunikative Beeinflussung, lassen aber noch keine Aussage darüber zu, ob die Zielgruppe die Botschaften oder Informationen möglicherweise ignoriert, nicht geglaubt oder direkt wieder vergessen hat. Auf der *indirekten Outcome-Ebene* kommen insbesondere quantitative Befragungen (z. B. per Smartphone-App, Paper-Pencil-Befragung), Fokusgruppen (z. B. im Rahmen einer Bürger:innenkonferenz), Beobachtungen (z. B. Wegeverlauf von Teilnehmer:innen bei einer Museumsausstellung) oder halbstandardisierte Interviews (z. B. mit Lehrer:innen oder Schüler:innen) zum Einsatz. Hier geht es dann um die Frage, ob und inwiefern sich Veränderungen von Wahrnehmungen, Wissen, Einstellungen, Emotionen, Verhaltensabsichten oder Verhaltensweisen zeitigen. Neben solchen formalen Methoden können auch informelle Gespräche mit Teilnehmer:innen eingesetzt werden, um Feedback einzuholen (Grand und Sardo 2017; vgl. auch Wollmann, Birkenhäger und Bastge in diesem Band).

3.2 Betriebswirtschaftliche Methoden

Der Einsatz von Methoden und Bewertungssystemen aus der Betriebswirtschaftslehre eignet sich insbesondere für die Evaluation der *Input-, internen Output- und der Impact-Ebene* der Wissenschaftskommunikation (Raupp 2017). Denn Maßnahmen der Wissenschaftskommunikation sollten nicht nur im Hinblick auf die beabsichtigte Wirkung einer Botschaft oder Kampagne bewertet werden, sondern auch im Hinblick auf die Frage, ob Aufwand und Ertrag in einem angemessenen Verhältnis stehen. Auf der *Input-Ebene* kann bspw. mithilfe von Budgetanalysen ermittelt werden (Zerfaß und Volk 2019, S. 84 ff.), wie viele personelle Ressourcen sowie Sach- oder Reisekosten in die Konzeption einer Veranstaltung geflossen sind. Auf der *internen Output-Ebene* geht es um die interne Effizienz bei der Produktion von Maßnahmen der Wissenschaftskommunikation: Hier können Prozessanalysen Aufschluss darüber geben (Zerfaß und Volk 2019, S. 41 ff.), inwiefern die Beteiligten zufrieden mit der Effizienz und Qualität von Prozessabläufen waren oder inwiefern Deadlines und Budgets eingehalten wurden. In Bezug auf die Erstellung von Content für Social Media oder Pressearbeit lässt sich bspw. analysieren, wie viele Korrekturschleifen es gab, wie hoch die Fehlerquote oder die Reaktionszeit auf Kommentare war (Raupp und Osterheider 2019). Auf der *Impact-Ebene* lassen sich für die Evaluation der langfristigen Beiträge von Wissenschaftskommunikation nur wenige der betriebswirtschaftlichen Methoden adaptieren, da es hier nicht oder selten um ökonomische Zielgrößen wie Umsatz oder Gewinn geht, sondern vielmehr um

gesellschaftliche (z. B. erzieherische, ökologische, etc.) Wirkungen. Hier lassen sich etwa Reputationsanalysen (z. B. in Bezug auf die Third Mission von Hochschulen) oder Akzeptanz- und Vertrauensanalysen (z. B. in das Wissenschaftssystem oder in wissenschaftliche Innovationen) mithilfe von Längsschnittdesigns durchführen. Darüber hinaus lässt sich bspw. auswerten, wie viele Projekt-/Förderabschlüsse (z. B. eingeworbene Drittmittel) oder neue Netzwerke bzw. Kooperationen realisiert wurden.

3.3 Kombination von Methoden

Da die Evaluation von Projekten oder Maßnahmen der Wissenschaftskommunikation entlang der Input–Output-Outcome-Impact-Stufen den Einsatz verschiedener Methoden erfordert, werden unterschiedliche Methoden in Abhängigkeit vom Zweck und Erkenntnisinteresse der Evaluation oft miteinander kombiniert. Tab. 1 gibt einen Überblick über typische Methoden und exemplarische Messgrößen entlang der vier Wirkungsstufen (in Anlehnung an Volk und Zerfaß 2022; Raupp und Osterheider 2019). Die Zusammenstellung ist nicht erschöpfend und einige Methoden können auf mehreren Wirkungsstufen zum Einsatz kommen. In der Regel gibt es keine Goldstandards für die Definition und Operationalisierung der hier dargestellten Messgrößen, sondern verschiedene Möglichkeiten und Herangehensweisen. Wenn Kennzahlen unterschiedlich definiert und erhoben werden, ist es schwierig, sinnvolle Vergleiche bzw. «Benchmarks» zwischen Projekten, Maßnahmen oder mit anderen Organisationen anzustellen. Daher ist es über verschiedene Projektteams hinweg bzw. innerhalb wissenschaftlicher Institutionen wichtig, einheitlich und konsistent zu definieren, ob es sich beispielsweise bei „visits" um „unique visits", „new visitors" oder „returning visitors" handelt, und identische Messzeiträume festzulegen (etwa pro Tag, Woche, Monat etc.). Auch für die Bestimmung von Engagement-Raten gibt es unterschiedliche Ansätze, Interaktionen auf Social Media (z. B. Likes, Comments, Retweets) in Beziehung zu den User:innen (z. B. Anzahl Follower, Abonnent:innen) zu setzen. Noch komplexer wird es, wenn etwa Reputation oder Glaubwürdigkeit der Wissenschaftskommunikation erhoben wird, denn auch hier liegen verschiedene Operationalisierungsansätze in der Forschung vor (vgl. Mede 2022). Dass es insgesamt bisher wenige standardisierte Kennzahlen gibt, auch im Feld der Strategischen Kommunikation, hat verschiedene Ursachen (z. B. Buhmann et al. 2019) und ist u. a. auf Marktdynamiken zurückzuführen. Für die Auswahl von geeigneten Evaluationsmethoden, Messgrößen und Kennzahlen für die Wissenschaftskommunikation gibt es also keine einheitliche Schablone; vielmehr sollte die

Tab. 1 Exemplarische Methoden und Messgrößen für die Evaluation von Wissenschaftskommunikation

Stufen	Methoden	Messgrößen
Input	Kostenerfassung Prozesskostenrechnung	Personaleinsatz Finanzaufwand
Interner Output	Budgetanalyse Prozessanalyse UX Research/Touchpoint-Analyse	Budgettreue, Prozessqualität Durchlaufzeiten, Fehlerquoten Verständlichkeit, Usability Anzahl Aktivitäten (z. B. Posts, Pressemitteilungen)
Externer Output	Inhaltsanalyse, Medienresonanzanalyse Nutzungsdatenanalyse (Webtracking) Eye Tracking	Reichweite, Share of Voice Tonalität der Berichterstattung Visits, Page Impressions Blickverlauf
Direkter Outcome	Beobachtung Befragung Social Media Analytics, Sentiment-Analyse Experimente, physiologische Messung	Anzahl Teilnehmende, Follower, Fans Recall, Recognition Likes, Shares, Comments Hautwiderstand
Indirekter Outcome	Befragung Fokusgruppen Beobachtung (z. B. Wegeverlauf) Feedbackmethoden Experimente, physiologische Messung	Interesse, Lernen, Wissen Engagement mit Wissenschaft Emotionen Verhaltensabsicht Weiterempfehlungsbereitschaft
Impact	Akzeptanzanalyse Vertrauensanalyse Reputationsbewertung	Gesellschaftliche Akzeptanz (z. B. in KI) Vertrauen (z. B. in Wissenschaft) Umsatz (z. B. Museum) Projekt-/Förderabschlüsse

Auswahl jeweils situations- und kontextspezifisch entlang der Ziele und Motive von Evaluationen erfolgen (Fu et al. 2016; Scheuerle 2020). Die Entscheidung, welche Kennzahlen und wie aufwendig diese Kennzahlen erhoben werden, hängt dabei auch von den verfügbaren Ressourcen ab. Letztlich kommt es bei der Definition von Kennzahlen darauf an, dass sie sich möglichst eng an der aktuellen Diskussion in der Forschung und Wissenschaftskommunikations-Community orientieren und idealerweise auch stetig weiterentwickelt werden, beispielsweise vor dem Hintergrund neuerer Forschungsergebnisse.

Für die Evaluationspraxis der Wissenschaftskommunikation bietet es sich an, die verschiedenen Methoden und Kennzahlen, die in einem Projekt oder Team genutzt werden, in einem (virtuellen) Methodenhandbuch – also einer eigenen *Toolbox* – zusammenzustellen. Eine solche Toolbox sollte idealerweise Beschreibungen, Vorlagen und Literaturempfehlungen zu den einzelnen Methoden sowie Kennzahlensteckbriefe beinhalten (z. B. mit einheitlichen Messgrößen, Messzeiträumen) (Zerfaß und Volk 2019). Verantwortliche für Wissenschaftskommunikation können sich dann mit den Voraussetzungen und Anwendungen spezifischer Methoden vertraut machen und geeignete Methoden für die eigene Fragestellung aussuchen. Eine zentrale Voraussetzung für empirisch aussagekräftige Messungen besteht darin, dass ausreichende Ressourcen und Zeit für Evaluationen zur Verfügung stehen und Wissenschaftskommunikator:innen solide Kenntnisse sozialwissenschaftlicher Methoden mitbringen (Jensen 2015; King et al. 2015). Daher sollte genügend Budget für Evaluationen schon im Vorfeld (etwa in Anträgen) einkalkuliert werden. Zum anderen sollten auch Methodenkompetenzen, die für die Konzeption von Evaluationsdesigns und Interpretation von Ergebnissen unerlässlich sind, systematisch aufgebaut und regelmäßig weiterentwickelt werden (Jensen und Gerber 2020; Pellegrini 2021; Ziegler et al. 2021).

4 Ausblick und Herausforderungen

Bislang steckt eine systematische Zieldefinition und Evaluation der Wissenschaftskommunikation noch weitestgehend in den Kinderschuhen und wird selten strukturiert angewandt (z. B. Jensen 2015; Weingart und Joubert 2019; Ziegler et al. 2021). Der Fokus liegt eher auf quantitativen Indikatoren und positiven Ergebnissen, obschon auch nicht intendierte, negative oder fehlende Effekte wichtige Erkenntnisse für die Wissenschaftskommunikation bereithalten (Jensen 2015). Oft werden einfach messbare Auszählungen wie etwa die Anzahl an Besucher:innen, Likes, Shares, Kommentare ausgewiesen oder lediglich Selbstauskünfte gesammelt (Weingart und Joubert 2019). Aufwendigere Verfahren wie Besucher:innenbefragungen oder -beobachtungen kommen zwar zum Einsatz, basieren aber teilweise auf wenig aussagekräftigen methodischen Designs (z. B. keine Vorher-Nachher-Befragung, Fehlen von validen und reliablen Instrumenten) und liefern teilweise verzerrte Datengrundlagen (Jensen 2014; Phillips et al. 2018). Dafür gibt es verschiedene Gründe, etwa das Fehlen von strukturellen und technischen Voraussetzungen, Ressourcen und methodischen Kompetenzen bei den Verantwortlichen oder einheitlichen und standardisierten Messgrößen (Fu et al. 2016; Jensen 2015).

Wenn aber evidenzbasierte Wissenschaftskommunikation in Zukunft an Bedeutung gewinnen wird (Jensen und Gerber 2020) und insbesondere in Deutschland durch die BMBF-Initiativen zur Förderung der Wissenschaftskommunikation zunehmend Steuergelder in ihre Professionalisierung investiert werden (BMBF 2019), wird eine systematische und wissenschaftlich robuste Evaluation an Relevanz gewinnen. Dabei sollte weniger die retrospektive Erfolgskontrolle im Sinne einer summativen Evaluation im Vordergrund stehen, sondern vielmehr der Versuch, durch kontinuierliche Evaluationen Lernprozesse zu ermöglichen und damit die Basis für evidenzbasierte Wissenschaftskommunikation zu schaffen (Raupp und Osterheider 2019; Ziegler et al. 2021). Denn das bildet letztlich die Grundlage dafür, Wissenschaftskommunikation datengestützt und empirisch fundiert einzusetzen – statt auf Bauchgefühl, Intuition und Erfahrungen zu hören. Gleichzeitig sollten sich Praktiker:innen in Wissenschaftskommunikation über die Voraussetzungen und Limitationen von Evaluationen bewusst sein und in Bezug auf die eingesetzten Methoden und Kennzahlen größtmögliche Transparenz herstellen (Volk und Zerfaß 2022). Ansatzpunkte hierfür bieten die Beiträge zu den Grundlagen, Voraussetzungen, Gütekriterien und Anwendungsbeispielen unterschiedlicher Methoden in diesem Band; ferner können Wissenschaftskommunikator:innen in der How-To-Reihe *Wisskomm evaluieren* der Impact Unit von Wissenschaft im Dialog (https://impactunit.de/tools/) oder im Leitfaden *Kommunikationscontrolling* des Bundesverbands Hochschulkommunikation Hilfestellung finden (Scheuerle et al. 2017). Für die Weiterentwicklung ihrer Evaluationspraktiken sollten Praktiker:innen schließlich auch regelmäßig die eigenen *Evaluationen evaluieren*, d. h. die jeweils eingesetzten Methoden, Zielgrößen, Kennzahlen oder Messzeiträume kritisch überprüfen und vor dem Hintergrund neuer Entwicklungen oder Standards ggf. aktualisieren (Jensen 2014).

In Zukunft wird insbesondere die Echtzeit-Evaluation von sozialen und alternativen Medien und neuen Gatekeepern (z. B. Influencern) in der Wissenschaftskommunikation an Bedeutung gewinnen, etwa um potenzielle Risiken (z. B. Desinformation, Anfeindungen gegenüber Wissenschaftler:innen) schnell erkennen und um aufkommenden Krisen kommunikativ begegnen zu können. Technologische Entwicklungen im Bereich der künstlichen Intelligenz und Automatisierung bieten hier neue Möglichkeiten, große unstrukturierte Datensätze (z. B. Sentiments von Twitter-Diskussionen) automatisiert und damit zeit- und ressourcenschonend in Echtzeit auszuwerten (Jensen 2015; Volk und Buhmann 2023). Die Verknüpfung von großen Datenmengen aus der Nutzungsdatenanalyse mit soziodemografischen und persönlichen Merkmalen von Zielgruppen (z. B. Alter, Kanalpräferenzen, Lebensstil, etc.) eröffnet für die Wissenschaftskommunikation zudem neue Möglichkeiten, ehemals breit definierte Zielgruppen

in kleinteilige Zielgruppen mit ähnlichen Profileigenschaften zu segmentieren („Profiling") und diese mit passend zugeschnittenen Kommunikationsbotschaften zu adressieren („Microtargeting"). Dies wirft natürlich ethische Fragestellungen (z. B. zu Data Privacy, gesellschaftlicher Polarisierung) auf, die es in Wissenschaftskommunikationspraxis und -forschung zu diskutieren gilt.

Aus Perspektive der Wissenschaftskommunikationsforschung stellen sich in Zukunft einerseits konzeptionelle Fragen in Bezug auf die Diskussion, welche generischen Ziele in der Wissenschaftskommunikationsforschung verfolgt und wie diese trennscharf definiert werden können (z. B. Metcalfe 2019; Ziegler et al. 2021). Insbesondere bedarf es einer Konzeptualisierung, welche langfristigen, gesellschaftlichen Beiträge auf der Impact-Ebene geschaffen werden (Fogg-Rogers et al. 2015; Weitkamp 2015). Andererseits lassen sich methodische Fragen in Bezug auf die empirische Messung von Konstrukten wie etwa Vertrauen in die Wissenschaft erörtern. Ferner ergeben sich empirische Fragen zu den Praktiken und Barrieren der Evaluation von Wissenschaftskommunikation. Dabei ließe sich eine Reihe von Fallstricken erforschen – etwa Overpromising bei der Selbstevaluation, Satisficing bei der Wahl von Evaluationsmethoden, Kennzahlenfixierung bei der Ergebnisinterpretation oder strategische Blindheit bei der Ableitung von Handlungsbedarf (Fischhoff 2019; Volk und Buhmann 2019).

Literatur

AMEC (2016) AMEC Integrated Evaluation Framework. https://amecorg.com/amecframework/de/framework/interactive-framework/. Zugegriffen: 1. März 2021

Besley JC, Dudo A (2022) Strategic Science Communication: A Guide to Setting the Right Objectives for More Effective Public Engagement. Johns Hopkins University Press, Baltimore

BMBF (2019) Grundsatzpapier des Bundesministeriums für Bildung und Forschung zur Wissenschaftskommunikation. https://www.bmbf.de/upload_filestore/pub/Grundsatzpapier_zur_Wissenschaftskommunikation.pdf. Zugegriffen: 11. Febr. 2022

Buhmann A, Likely F (2018) Evaluation and measurement. In: Heath RL, Johansen W (Hrsg) The international encyclopedia of strategic communication. Wiley-Blackwell, Malden, S 625–640

Buhmann A, Volk SC (2022) Measurement and evaluation: framework, methods, and critique. In: Falkheimer J, Heide M (Hrsg) Handbook of strategic communication. Edward Elgar, Cheltenham, S 474–488. https://doi.org/10.4337/9781800379893.00039

Buhmann A, Macnamara J, Zerfass A (2019) Reviewing the 'march to standards' in public relations: a comparative analysis of four seminal measurement and evaluation initiatives. Public Relat Rev 45(4):101825. https://doi.org/10.1016/j.pubrev.2019.101825

Clark H, Taplin D (2012) Theory of change basics: a primer on theory of change. Actknowledge, New York

DPRG, ICV (2011) Positionspapier Kommunikationscontrolling. DPRG/ICV, Bonn, Gauting

Fischhoff B (2019) Evaluating science communication. Proceedings of the National Academy of Sciences of the USA 116(16):7670–7675. https://doi.org/10.1073/pnas.1805863115

Fogg-Rogers L, Sardo AM, Grand A (2015) Beyond dissemination – science communication as impact. J Sci Commun 14(03):1–7. https://doi.org/10.22323/2.14030301

Frechtling JA (2015) Logic modeling methods in program evaluation. Wiley, New York

Fu AC, Kannan A, Shavelson RJ, Peterson L, Kurpius A (2016) Room for rigor: designs and methods in informal science education evaluation. Visitor Studies 19(1):12–38. https://doi.org/10.1080/10645578.2016.1144025

Funnell S, Rogers P (2011) Purposeful program theory: effective use of theory of change and logic models. Jossey-Bass, CA

Grand A, Sardo AM (2017) What works in the field? Evaluating informal science events. Front Commun 2(22). https://doi.org/10.3389/fcomm.2017.00022

Jensen E (2014) The problems with science communication evaluation. J Sci Commun 13:C04. https://doi.org/10.22323/2.13010304

Jensen E (2015) Evaluating impact and quality of experience in the 21st century: using technology to narrow the gap between science communication research and practice. J Sci Commun 14(03). https://doi.org/710.22323/2.14030305

Jensen E, Gerber A (2020) Evidence-Based science communication. Front Commun 4:78. https://doi.org/10.3389/fcomm.2019.00078

King H, Svalastog AL, Hobson M, Robinson A, Clipson H (2015) Highlighting the value of evidence-based evaluation: pushing back on demands for ‘impact’. J Sci Commun 14(02). https://doi.org/10.22323/2.14020202

Knowlton L, Phillips C (2013) The logic models guidebook: better strategies for great results, 2. Aufl. Sage, London

Macnamara J (2018) Evaluating public communication: new models, standards, and best practice. Routledge, London

Macnamara J, Gregory A (2018) Expanding evaluation to progress strategic communication: beyond message tracking to open listening. Int J Strateg Commun 12(4):469–486. https://doi.org/10.1080/1553118X.2018.1450255

Mede NG (2022) Charakteristika der Forschung zu Wirkungen digitaler Wissenschaftskommunikation: Ein Systematic Review der Fachliteratur. Wissenschaftspolitik im Dialog 17(47). Berlin-Brandenburgische Akademie der Wissenschaften, S 37–82

Metcalfe J (2019) Comparing science communication theory with practice: An assessment and critique using Australian data. Public Underst Sci 28(4):382–400. https://doi.org/10.1177/0963662518821022

Pellegrini G (2021) Evaluating science communication. Concepts and tools for realistic assessment. In Bucchi M, Trench B (Hrsg.) Routledge Handbook of Public Communication of Science and Technology, 3. Aufl. Routledge, London, S 305–322. https://doi.org/10.4324/9781003039242

Phillips T, Porticella N, Constas M, Bonney R (2018) A framework for articulating and measuring individual learning outcomes from participation in citizen science. Citizen Sci: Theor Pract 3(2). https://doi.org/10.5334/cstp.126

Raupp J (2017) Strategische Wissenschaftskommunikation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer VS, Wiesbaden, S 143–163

Raupp J, Osterheider A (2019) Evaluation von Hochschulkommunikation. In: Fähnrich B, Metag J, Post S, Schäfer MS (Hrsg) Forschungsfeld Hochschulkommunikation. Springer VS, Wiesbaden, S 181–205. https://doi.org/10.1007/978-3-658-22409-7_9
Raupp J, Vogelgesang J (2009) Medienresonanzanalyse. Springer VS, Wiesbaden
Rice R, Atkin C (2013) Public communication campaigns, 4. Aufl. Sage, Thousand Oaks, Calif
Rossi P, Lipsey M, Freeman H (2004) Evaluation: A systematic approach, 7. Aufl. Sage, Thousand Oaks, Calif
Scheuerle S (2020) Kommunikations-Controlling. In: Welpe IM, Stumpf-Wollersheim J, Folger, N Prenzel M (Hrsg), Leistungsbewertung in wissenschaftlichen Institutionen und Universitäten: Eine mehrdimensionale Perspektive. De Gruyter & Oldenbourg, Berlin, S 187–204
Scheuerle S, Flacke M, Jordan K, Kiechle D, Maas S, Meyer K, Müller-Detert U, Pernat N, Sommerfeld U, von Soosten C (2017) Leitfaden Kommunikations-Controlling. https://iqcc.jimdofree.com. Zugegriffen: 11. Febr. 2022
Volk SC (2016) A systematic review of 40 years of public relations evaluation and measurement research: looking into the past, the present, and future. Pub Relat Rev 42(5):962–977. https://doi.org/10.1016/j.pubrev.2016.07.003
Volk SC, Buhmann A (2019) New avenues in communication evaluation and measurement. Towards a research agenda for the 2020s. J Commun Manage 23(3):162–178. https://doi.org/10.1108/JCOM-08-2019-147
Volk SC, Buhmann A (2023) Evaluation and Measurement in the digital age: challenges and opportunities for corporate communication. In: Luoma-aho M, Badham A (Hrsg) Handbook of digital corporate communication. Edward Elgar, Cheltenham
Volk SC, Zerfaß A (2022) Kommunikationscontrolling und PR-Evaluation. In: Szyszka P, Fröhlich R, Röttger U (Hrsg) Handbuch der Public Relations, 4. Aufl. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-28149-6_51-2
Watson T, Noble P (2014) Evaluating public relations, 3. Aufl. Kogan Page, London
Weingart P, Joubert M (2019) The conflation of motives of science communication – causes, consequences, remedies. J Sci Commun 18(Y01). https://doi.org/10.22323/2.18030401
Weitkamp E (2015) Between ambition and evidence. J Sci Commun 14(E). https://doi.org/10.22323/2.14020501
Zerfaß A, Volk SC (2019) Toolbox Kommunikationsmanagement. Denkwerkzeuge und Methoden für die Steuerung der Unternehmenskommunikation. Springer Gabler, Wiesbaden
Ziegler R, Hedder IR, Fischer L (2021) Evaluation of science communication: current practices, challenges, and future implications. Front Commun 6:669744. https://doi.org/10.3389/fcomm.2021.669744

Sophia C. Volk ist Oberassistentin in der Abteilung Wissenschaftskommunikation am Institut für Kommunikationswissenschaft und Medienforschung der Universität Zürich. Sie forscht zu den gesellschaftlichen Wirkungen der Wissenschaftskommunikation, zu den Strukturen der Hochschulkommunikation und zur Kommunikation von Wissenschaftler:innen in der Öffentlichkeit und in Forschungsverbünden. In ihrer Forschung kombiniert sie quantitative und qualitative sowie international vergleichende Methoden.


cc
BY



Evaluationsstandards – Leitprinzipien von Evaluationen

Vanessa van den Bogaert

Zusammenfassung

Woran sind gute Evaluationen zu erkennen? Wie lassen sich praxistaugliche Evaluationsvorhaben auf hohem Niveau realisieren? Zur Beantwortung dieser Fragen werden im gleichnamigen Beitrag die Evaluationsstandards der *DeGEval* (Deutsche Gesellschaft für Evaluation) eingeführt. Dazu werden die 25 Einzelstandards, geordnet nach den vier Qualitätsdimensionen *Nützlichkeit*, *Durchführbarkeit*, *Fairness* und *Genauigkeit* detailliert beschrieben und erklärt.

1 Einleitung

Das Themen- und Forschungsfeld der Evaluation hat sich in den letzten Jahrzehnten zu einem interdisziplinären Arbeitsgebiet entwickelt, mit speziellen Methoden und Standards, Fachgesellschaften und Kongressen, Handbüchern und Zeitschriften (Westermann 2016). Dem Bedarf der Community nach möglichst allgemeingültigen Evaluationsstandards wurde in den USA bereits in den 1970er-Jahren mit dem sog. *Joint Committee on Standards for Educational Evaluation* nachgegangen (Sanders 2013). Zeitlich versetzt wurden auch in den D-A-CH-Ländern – von den jeweils dort ansässigen nationalen Evaluationsgesellschaften

V. van den Bogaert (✉)
Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland
E-Mail: vanessa.vandenbogaert@ruhr-uni-bochum.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_4

(*DeGEval* – Deutsche Gesellschaft für Evaluation; *fteval* – Österreichischen Plattform für Forschungs- und Technologiepolitikevaluation; *SEVAL* – Schweizerische Evaluationsgesellschaft) – Evaluationsstandards erarbeitet und verabschiedet. In diesem Beitrag liegt der Fokus daher auf den Evaluationsstandards, welche ganz allgemein als wünschbare Eigenschaften von Evaluationen gelesen werden können (Beywl 2019). Die Standards bieten Unterstützung bei der Klärung von Fragen nach dem Zweck, der Planung, Durchführung, Auswertung sowie Kommunikation während der Informationssammlung und der Berichtslegung von Evaluationen. Die Standards der drei o. g. Fachgesellschaften unterscheiden sich zwar hinsichtlich ihres inhaltlichen Aufbaus, sie geben jedoch alle einen allgemeingültigen Rahmen für professionelle Evaluationen und legen damit einen Grundstein für eine planvolle und allgemein akzeptable Vorgehensweise. Heute werden im deutschsprachigen Raum dazu vier Qualitätsdimensionen von Evaluationen aufgeführt: *Nützlichkeit, Durchführbarkeit, Fairness* und *Genauigkeit* (DeGEval 2017). Darüber hinaus bieten die Hinweise für Evaluationen in der Wissenschaftskommunikation der *Impact Unit* spezifische Unterstützung in dem Themenbereich dieses Bands. Das *Factsheet* Evaluationsstandards bietet eine konkrete Hilfestellung, indem die Evaluationsstandards der *DeGEval* für die Wissenschaftskommunikation gerahmt und erläutert werden.

2 Standards für Evaluation

Die Standards für Evaluationen können auch eine Richtschnur zur Bewertung von Evaluationsplänen und -berichten sein und sollten bewusst nicht als ein restriktives Regelwerk gelesen werden (Sanders 2013). Es ist wichtig anzuerkennen, dass nicht alle Standards immer vollumfänglich Berücksichtigung finden können. Die Standards sind zudem als Maximalstandards formuliert worden, dabei können die Einzelstandards teilweise sogar in Konkurrenz zueinander stehen (DeGEval 2017). Daher sollten Evaluator:innen und/oder damit betraute Personen immer eine angemessene sowie zweckdienliche Nutzung der Standards sicherstellen. Die im jeweiligen Kontext anwendbaren Standards müssen dazu zunächst identifiziert und bei der Planung und Durchführung der Evaluation berücksichtigt werden. Gleichermaßen sollte eine bewusste Nichterfüllung von (Einzel-)Standards offen und nachvollziehbar dokumentiert werden. Nachfolgend werden die Standards für Evaluationen der *DeGEval* benannt und anschließend kurz erläutert.

2.1 Nützlichkeit

Die Nützlichkeitsstandards beschreiben zunächst die Orientierung einer Evaluation an den Informationsbedürfnissen der Nutzer:innen. Die *DeGEval* hat dazu acht Teilstandards formuliert (DeGEval 2017).

2.1.1 Identifizierung der Beteiligten und Betroffenen

Standard N1

> „Die am Evaluationsgegenstand oder an der Evaluation Beteiligten sowie die von Evaluationsgegenstand oder Evaluation Betroffenen sollen vorab identifiziert werden, damit deren Interessen und Informationsbedürfnisse geklärt und so weit wie möglich bei der Anlage der Evaluation berücksichtigt werden können“ (DeGEval 2017, S. 34).

Zumeist gibt es eine Vielzahl an unterschiedlichen Stakeholder:innen und damit an Personen und/oder Personengruppen, die in ein Evaluationsvorhaben einbezogen werden sollten oder von ihm betroffen sein könnten (Sanders 2013). Entscheidend ist hier die Ermittlung der Informationsbedürfnisse, um das Evaluationsvorhaben entsprechend passgenau ausrichten zu können. Dabei sollte bereits zu Beginn festgelegt werden, welche Bedeutung den Informationswünschen der Beteiligten und Betroffenen beigemessen werden kann – und wie diese im Rahmen der zeitlichen und finanziellen Ressourcen realisierbar sind.

2.1.2 Klärung der Evaluationszwecke

Standard N2

> „Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen und die Evaluierenden einen klaren Arbeitsauftrag verfolgen können“ (DeGEval 2017, S. 35).

Die Festlegung von Evaluationszwecken ist eine notwendige Grundlage für die Planung und Durchführung von Evaluationsvorhaben. Evaluationen sollen Orientierung geben können. Im Sinne von formativen Evaluationen kann daher der Hauptzweck, bspw. in der Bereitstellung von Informationen zur weiterführenden Gestaltung des Evaluationsgegenstandes liegen (siehe auch Volk in diesem Band). Demgegenüber kann der Hauptzweck bei summativen Evaluationen in der Informationsbeschaffung für grundlegende Entscheidungen

liegen, bspw. um die Ausweitung oder Einstellung eines Programms bzw. einer Maßnahme mit empirischen Informationen zu unterstützen. Evaluationen können aber auch Diskussionen im öffentlichen, politischen oder wissenschaftlichen Raum anstoßen und stellen zu diesem Zweck systematisch gewonnene Erkenntnisse bereit (DeGEval 2017).

2.1.3 Kompetenz und Glaubwürdigkeit des Evaluators/der Evaluatorin

Standard N3

> „Wer Evaluationen durchführt, soll fachlich und methodisch kompetent sein, damit für die Evaluation und ihre Ergebnisse ein Höchstmaß an Glaubwürdigkeit und Akzeptanz erreicht wird" (DeGEval 2017, S. 36).

Sach- und Fachwissen, Ausbildung, technische Kompetenz, Erfahrung, Integrität von Evaluator:innen oder damit betrauten Teams stellen Eigenschaften dar, die eine Glaubwürdigkeitszuschreibung von Dritten beeinflussen. Da selten eine Person all den Anforderungen genügt, ist es immer empfehlenswert als Kollektiv gemeinsam über all diese Kompetenzen zu verfügen. Eine frühzeitige Überprüfung der Glaubwürdigkeit wird den Teams später dabei helfen die Nützlichkeit, Relevanz und Angemessenheit der Befunde bekräftigen zu können. Die *DeGEval* hat darüber hinaus auch sog. Evaluationskompetenzen für Evaluator:innen in Form von Anforderungsprofilen ausgearbeitet (detaillierte Ausführung hierzu s. DeGEval 2004b).

2.1.4 Auswahl und Umfang der Informationen

Standard N4

> „Auswahl und Umfang der erfassten Informationen sollen die adäquate Beantwortung der zu untersuchenden Fragestellungen zum Evaluationsgegenstand ermöglichen und die Informationsbedürfnisse der Auftraggebenden und weiterer Beteiligter und Betroffener berücksichtigen" (DeGEval 2017, S. 36).

Es ist in der Regel nicht möglich, alle potenziell verfügbaren Informationen zusammenzutragen und zu analysieren. Die gewonnenen Informationen müssen jedoch so umfangreich sein, dass alle wichtigen Dimensionen des zu evaluierenden Gegenstandsbereichs erfasst und dokumentiert werden können. Um auf Unwichtiges bewusst verzichten zu können, müssen daher Abwägungen

getroffen werden. Dazu müssen Fragestellungen herausgefiltert und gewichtet werden (Sanders 2013). Da sich Informationsbedürfnisse auch im Laufe der Zeit verändern können, sollte der Austausch über eine mögliche Aktualisierung regelmäßig gesucht werden (Sanders 2013). Themenbezogene Literatur, frühe Evaluationsberichte, Austausche in Netzwerken und moderierte Diskussionen können diesen Prozess unterstützen (DeGEval 2017).

2.1.5 Transparenz von Werthaltungen

Standard N5

> „Werthaltungen der Beteiligten und Betroffenen, die sich in deren Perspektiven und Annahmen manifestieren und einen Einfluss haben auf die Evaluation und Interpretation ihrer Ergebnisse, sollten transparent dokumentiert werden, um Evaluationsergebnisse besser einordnen zu können" (DeGEval 2017, S. 37).

Wie bereits im Einleitungskapitel herausgestellt wurde, ist das Bewerten immanenter Bestandteil von Evaluationen. Eine Bewertung beschreibt somit die Einschätzung einer Sache nach ihrer Nützlichkeit und ihrem allgemeinen Wert (Sanders 2013). Die Interpretation der gesammelten Informationen stellt den wichtigsten Punkt im Evaluationsprozess dar (DeGEval 2017). Alle gewonnenen Informationen – quantitativ, qualitativ, prozess- oder produktbezogen, formativ oder summativ – müssen daher anhand zuvor festgelegter, transparenter und nachvollziehbarer Kriterien interpretiert und bewertet werden. Für eine solche angemessene Ergebnisinterpretation muss dargelegt werden, wer die Werturteile vornimmt und welche Verfahren dazu Verwendung finden (Sanders 2013). Evaluator:innen oder damit betraute Personengruppen handeln jedoch nicht per se objektiv und wertfrei – vielmehr müssen die Bewertungskriterien vorab sorgfältig erarbeitet werden.

2.1.6 Vollständigkeit und Klarheit der Berichterstattung

Standard N6

> „Evaluationsberichte sollen alle wesentlichen Informationen zur Verfügung stellen und für ihre Adressatinnen und Adressaten verständlich und nachvollziehbar sein" (DeGEval 2017, S. 37 f.).

Zwischenergebnisse und Schlussberichte sollen allen Nutzer:innen dienlich sein, weshalb der Vermittlung der Evaluationsergebnisse in Form von Berichten besondere Aufmerksamkeit zu schenken ist. Es gilt zu antizipieren, zu welchen Zeitpunkten die spezifischen Informationen am besten genutzt werden können. Daher empfiehlt es sich, alle vorgesehenen Nutzer:innen der Berichte zu identifizieren und passende Ansätze für verschiedene Nutzer:innengruppen zu erarbeiten. Hier wird besonders deutlich, dass das Evaluationsteam zusammen mit den Auftraggeber:innen der Informationspflicht nachkommen sollte. Eine enge Abstimmung und ggf. schriftliche Vereinbarung können möglichen Interessenskonflikten bei der Identifikation von potenziellen Nutzer:innen entgegenwirken. Insbesondere in sog. Selbstevaluationsberichten sollten die wesentlichen Informationen (Beschreibung des Evaluationsgegenstandes, des Kontextes, der Ziele, der Fragestellungen, der Verfahren sowie der Befunde) in verständlicher Weise allen Beteiligten und Betroffenen zugänglich gemacht werden (DeGEval 2004a).

2.1.7 Rechtzeitigkeit der Evaluation

Standard N7

> „Evaluationsvorhaben sollen so rechtzeitig begonnen und abgeschlossen werden, dass ihre Ergebnisse in anstehende Entscheidungs-, Verbesserungs- oder sonstige Nutzungsprozesse einfließen können“ (DeGEval 2017, S. 38 f.).

Hier geht es vor allem um Zeitpläne, die die notwendigen Abstimmungsprozesse, Vorbereitungsphasen, bestenfalls sogar Pufferzeiten sowie Zeiten für eine sorgfältige Fehlerprüfung und Auswertung realistisch abbilden. Die *DeGEval* weist darauf hin, dass in vielen Fällen eine beträchtliche Vorlaufzeit eingeplant werden muss, da interne Abstimmungsprozesse (z. B. Stellungnahmen oder Mitzeichnungsverfahren) vor der Veröffentlichung eingehalten werden müssen. Rückmeldungsschleifen können auch so angesetzt werden, dass Nutzer:innen bspw. bereits durch Zwischenberichte handlungsrelevante Informationen beziehen können. (DeGEval 2017).

2.1.8 Nutzung und Nutzen der Evaluation

Standard N8

> „Planung, Durchführung und Berichterstattung einer Evaluation sollen die Beteiligten und Betroffenen dazu ermuntern, die Evaluation mitzutragen und ihre Ergebnisse zu nutzen" (DeGEval 2017, S. 38 f.).

Der mögliche Nutzen von Evaluationen ist nicht immer allen Beteiligten und Betroffenen von Anfang an ersichtlich. Daher empfiehlt es sich, bei der Kommunikation über das Vorhaben auch zu klären, in welcher Weise die Ergebnisse genutzt werden können. Besonders bei der Berichtslegung empfiehlt es sich, genügend Ressourcen einzuplanen, um über den Bericht und die erzielten Ergebnisse in den Austausch treten zu können.

2.2 Durchführbarkeit

Die Durchführbarkeitsstandards sollen den Rahmen für realistische, gut durchdachte sowie kostenbewusste Evaluationsvorhaben geben. Dazu zählen neben vertraglichen Vereinbarungen vor allem die Auswahl der Datenquellen, die Entscheidung für Methoden der Datensammlung, aber auch die Speicherung und Aufbereitung und Analyse der Daten sowie die Ergebnisaufbereitung. Gleichzeitig sollten Alternativen entwickelt werden, die einem ein rasches Ausweichen bei unvorhergesehenen Problemen ermöglichen. Die *DeGEval* hat dazu drei Durchführbarkeitsstandards formuliert, die im Folgenden erläutert werden.

2.2.1 Angemessene Verfahren

Standard D1

> „Evaluationsverfahren, einschließlich der Verfahren zur Beschaffung notwendiger Informationen, sollen so gewählt werden, dass einerseits die Evaluation professionell und den Erfordernissen entsprechend umgesetzt wird und andererseits der Aufwand für die Beteiligten und Betroffenen in einem adäquaten Verhältnis zum intendierten Nutzen der Evaluation gehalten wird" (DeGEval 2017, S. 39 f.).

Evaluationen bedeuten für alle beteiligten Personen – auch außerhalb des Evaluationsteams – einen Mehraufwand. Daher sollten möglichst wohldurchdachte Abwägungen zwischen den Anforderungen wissenschaftlicher

Güte und minimaler Störung der beteiligten Personen und/oder Organisationen getroffen werden. Die so erarbeiteten Vor- und Nachteile sowie die Angemessenheit und Aussagekraft der gewählten Herangehensweisen müssen dabei nachvollziehbar begründet werden (DeGEval 2017).

2.2.2 Diplomatisches Vorgehen

Standard D2

> „Evaluationen sollen so geplant und durchgeführt werden, dass eine möglichst hohe Akzeptanz der verschiedenen Beteiligten und Betroffenen in Bezug auf Vorgehen und Ergebnisse der Evaluation erreicht werden kann" (DeGEval 2017, S. 40).

Evaluationen können durchaus politische Tragweite haben, wenn z. B. Ergebnisse genutzt werden, um Entscheidungen zur Umverteilung von Ressourcen herbeizuführen. Verschiedene Standpunkte, aber auch Bedenken einzelner Interessensgruppen sind ernst zu nehmen. Es gilt das Vertrauen der Beteiligten und Betroffenen zu gewinnen und so versichern zu können, dass die Evaluation auf faire Weise erfolgt, d. h. ohne einzelnen Gruppen womöglich Vorteile zu verschaffen.

2.2.3 Effizienz von Evaluation

Standard D3

> „Der Aufwand für Evaluation soll in einem angemessenen Verhältnis zum Nutzen der Evaluation stehen" (DeGEval 2017, S. 40).

Das Aufwand-Nutzen-Verhältnis sollte bereits in der Planung von Evaluationen reflektiert und nachvollziehbar dargestellt werden. Ziel ist es, herauszustellen, welcher Aufwand entsteht und welcher Nutzen erwartet wird. Dabei kann der Nutzen als Wirkung der Evaluation verstanden werden, was direkte und indirekte, intendierte und nicht-intendierte Wirkungen umfasst (Sanders 2013).

2.3 Fairness

Die Fairnessstandards thematisieren den respektvollen und fairen Umgang mit allen beteiligten und betroffenen Personen und Gruppen (DeGEval 2017). Die *DeGEval* hat dazu die vier Standards formuliert.

2.3.1 Formale Vereinbarungen

Standard F1

> „Die Rechte und Pflichten der an einer Evaluation beteiligten Parteien (was, wie, von wem, wann getan werden soll und darf) sollen schriftlich festgehalten werden“ (DeGEval 2017, S. 41).

Eine schriftliche Vereinbarung stellt bestenfalls eine gemeinsame Referenz dar und hilft so dabei, dass sich alle Verantwortlichen an die vereinbarten und gemeinsam getragenen Bedingungen halten können. Mit formalen Vereinbarungen kann zumindest versucht werden, mögliche Missverständnisse frühzeitig zu vermeiden oder diese rechtzeitig zu bereinigen sowie Rechte und Pflichten transparent auszuhandeln (DeGEval 2017). Meist handelt es sich um folgende Bereiche: Finanzen, Zeit, Methodik, Veröffentlichungsrechte sowie mitwirkende Personen und Gruppen (DeGEval 2017).

2.3.2 Schutz individueller Rechte

Standard F2

> „Evaluationen sollen so geplant und durchgeführt werden, dass Rechte, Sicherheit und Würde der in eine Evaluation einbezogenen Personen geschützt sind“ (DeGEval 2017, S. 41 f.).

Der Schutz individueller Rechte steht über dem Interesse an Informationen (DeGEval 2017). Evaluationsteams sollten daher die gesetzlichen, behördlichen und/oder organisationsbezogenen Bestimmungen zum Datenschutz kennen und einhalten.

2.3.3 Umfassende und faire Prüfung

Standard F3

> „Evaluationen sollen die Stärken und die Schwächen des Evaluationsgegenstandes möglichst fair und umfassend prüfen und darstellen“ (DeGEval 2017, S. 42).

Stärken und Schwächen sollten gleichermaßen offengelegt werden. Dazu ist die Identifikation von unterschiedlichen Sichtweisen verschiedener Stakeholder:innen

unerlässlich. Ziel ist es, umfassende und gleichzeitig ausgewogene Evaluationen sicherzustellen (DeGEval 2017).

2.3.4 Unparteiische Durchführung und Berichterstattung

Standard F4

> „Die Evaluation soll unterschiedliche Sichtweisen von Beteiligten und Betroffenen auf Gegenstand und Ergebnisse der Evaluation beachten. Der gesamte Evaluationsprozess sowie die Evaluationsberichte sollen die unparteiische Position der Evaluierenden erkennen lassen" (DeGEval 2017, S. 42 f.).

Die vielfältigen Sichtweisen können durchaus verschieden sein. Evaluator:innen sollten daher eine möglichst unparteiische Position beziehen. Dies kann beispielsweise durch eine regelmäßige und planvolle Reflexion gelingen (DeGEval 2017).

2.3.5 Offenlegung von Ergebnissen und Berichten

Standard F5

> „Evaluationsergebnisse und -berichte sollen allen Beteiligten und Betroffenen soweit wie möglich zugänglich gemacht werden" (DeGEval 2017, S. 43).

Die Offenlegung von Berichten sollte bereits zu Beginn von Evaluationen sorgsam thematisiert werden. Die Evaluationsergebnisse und -berichte sollten i. d. R. allen Beteiligten und Betroffenen zugänglich gemacht werden, jedoch ist eine vollständige Offenlegung aus unterschiedlichsten Gründen nicht immer möglich (DeGEval 2017). Die *DeGEval* empfiehlt, bereits in der Planungsphase über die Art und den Umfang der Offenlegung der Ergebnisse zu entscheiden. Der Spielraum für eine kontextspezifische Lösung liegt dabei zwischen einer vollumfänglichen Veröffentlichung und einer begründeten Auswahl von (Teil-) Ergebnissen.

2.4 Genauigkeit

Die Genauigkeitsstandards thematisieren vor allem die wissenschaftlichen Gütekriterien und setzen die Gültigkeit und Nachvollziehbarkeit der Ergebnisse und Schlussfolgerungen in den Fokus (DeGEval 2017). Die *DeGEval* hat dazu acht Standards formuliert.

2.4.1 Beschreibung des Evaluationsgegenstandes

Standard G1

> „Sowohl das Konzept des Evaluationsgegenstand[e]s als auch seine Umsetzung sollen genau und umfassend beschrieben und dokumentiert werden" (DeGEval 2017, S. 44).

Herauszuarbeiten, welchen Geltungsbereich Evaluationen haben, bedeutet eine lückenlose und sorgfältige Beschreibung des Evaluationsgegenstandes. Dabei soll eine solche Beschreibung die beteiligten Akteur:innen, Ziele und Zwecke, Rahmenbedingungen und Strukturen sowie die bereitgestellten Ressourcen umfassen. Gleichzeitig sind auch die notwendigen Änderungen und Anpassungen wichtige Bestandteile einer solchen lückenlosen Beschreibung. Ziel ist, eine fehlerhafte Interpretation zu vermeiden und die wesentlichen Bedingungen für den möglichen Erkenntnistransfer transparent aufzeigen zu können (DeGEval 2017).

2.4.2 Kontextanalyse

Standard G2

> „Der Kontext des Evaluationsgegenstandes soll ausreichend umfassend und detailliert analysiert sowie bei der Interpretation von Ergebnissen berücksichtigt werden" (DeGEval 2017, S. 44 f.).

Um die Ergebnisse von Evaluationen hinsichtlich ihrer Gültigkeit und Übertragbarkeit einschätzen zu können, sind Erkenntnisse über mögliche Bedingungsfaktoren notwendig. Dies betrifft politische, ökonomische, soziale, technologische oder ökologische Rahmenbedingungen, die bei Evaluationen nicht ausgeblendet werden sollten (DeGEval 2017). Um mögliche Faktoren zu identifizieren, die einen maßgeblichen Einfluss auf den Evaluationsgegenstand und/oder die Ergebnisse haben können, sind Kontextanalysen ein wichtiger Bestandteil von Evaluationsvorhaben, denen genügend Aufmerksamkeit geschenkt werden sollte.

2.4.3 Beschreibung von Zwecken und Vorgehen

Standard G3

> „Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen so genau dokumentiert und beschrieben werden, dass sie nachvollzogen und beurteilt werden können“ (DeGEval 2017, S. 45).

Eine genaue und vollständige Dokumentation beinhaltet die Beschreibung von Zwecken, Fragestellungen, Zeitplänen, Vorgehensweisen und Methoden sowie mögliche Abweichungen.

2.4.4 Angabe von Informationsquellen

Standard G4

> „Die im Rahmen einer Evaluation genutzten Informationsquellen sollen hinreichend genau dokumentiert werden, damit die Verlässlichkeit und Angemessenheit der Informationen eingeschätzt werden können“ (DeGEval 2017, S. 45 f.).

Um die Qualität und Glaubwürdigkeit von Evaluationen einschätzen zu können, müssen die genutzten Informationsquellen lückenlos beschrieben werden (DeGEval 2017).

2.4.5 Valide und reliable Informationen

Standard G5

> „Erhebungsverfahren und Datenquellen sollen so gewählt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. Die fachlichen Maßstäbe sollen sich an den Gütekriterien der empirischen Forschung orientieren“ (DeGEval 2017, S. 46).

Die Methoden und Verfahren, die bei Evaluationen genutzt werden können, sind immer kontextabhängig. In diesem Band ist eine ganze Fülle von methodischen Herangehensweisen gesammelt worden. Unabhängig von den individuell gewählten methodischen Zugängen steht immer der Anspruch an eine möglichst hohe Validität der Schlussfolgerungen an erster Stelle. Dies betrifft die Güte und die Glaubwürdigkeit der Schlüsse, die aus den Ergebnissen des

Informationsgewinnungsprozesses gezogen werden können (Sanders 2013). Ziel ist es, konsistente, reproduzierbare, intersubjektiv nachvollziehbare, zulässige und valide Informationen bereitzustellen (DeGEval 2017).

2.4.6 Systematische Fehlerprüfung

Standard G6

> „Die in einer Evaluation gesammelten, aufbereiteten, analysierten und präsentierten Informationen sollen systematisch auf Fehler geprüft werden" (DeGEval 2017, S. 46 f.).

Empirisches Arbeiten, d. h. das Sammeln, Verarbeiten, Analysieren und Dokumentieren von Daten ist ein Prozess, der durchaus Fehlermöglichkeiten birgt (Sanders 2013). Mit einer systematischen Informationsprüfung soll sichergestellt werden, dass die Informationen so weit wie möglich fehlerfrei und abgesichert sind (Sanders 2013). Darum empfiehlt es sich, systematisch nach Fehlern zu forschen und das Thema der Qualitätssicherung ernst zu nehmen.

2.4.7 Angemessene Analyse qualitativer und quantitativer Informationen

Standard G7

> „Qualitative und quantitative Informationen einer Evaluation sollen nach fachlichen Maßstäben angemessen und systematisch analysiert werden, damit die Fragestellungen der Evaluation beantwortet werden können" (DeGEval 2017, S. 47 f.).

Die Analyseverfahren sind so auszuwählen, dass sie für die Fragestellungen der Evaluationen sowie für die Art der Daten angemessen sind. Dabei können sich qualitative und quantitative Zugänge gegenseitig ergänzen und die Interpretationen gemeinsam tragen (Sanders 2013).

2.4.8 Begründete Bewertungen und Schlussfolgerungen

Standard G8

> „Die in einer Evaluation getroffenen wertenden Aussagen sollen auf expliziten Kriterien und Zielwerten basieren. Schlussfolgerungen sollen ausdrücklich und auf Grundlage der erhobenen und analysierten Daten begründet werden, damit sie nachvollzogen und beurteilt werden können" (DeGEval 2017, S. 48).

An Schlussfolgerungen wird der Anspruch gestellt, sowohl die Fragestellungen wahrheitsgemäß zu beantworten, als auch das Vorgehen und die Ergebnisse wahrheitsgetreu widerzuspiegeln (Sanders 2013). Mögliche Limitationen der Evaluationsverfahren und der gewonnenen Datengrundlage sollten gleichermaßen thematisiert werden. Schlussfolgerungen sollten sich daher niemals auf unzureichende und ungenaue Informationen stützen, auch wenn der Wunsch nach der Beantwortung der Fragestellungen noch so groß erscheint (DeGEval 2017).

2.4.9 Meta-Evaluation

Standard G9

> „Meta-Evaluationen evaluieren Evaluationen. Um dies zu ermöglichen, sollen Evaluationen in geeigneter Form dokumentiert, archiviert und soweit wie möglich zugänglich gemacht werden“ (DeGEval 2017, S. 48 f.).

Auch wenn Meta-Evaluationen nicht im Fokus dieses Beitrags stehen, sind Systematisierungen von Evaluationen nur möglich, wenn einzelne Evaluationsvorhaben den Standards genügen. Daher sei an dieser Stelle angemerkt, dass eine angemessene Dokumentation und Berichtslegung Meta-Evaluationen, Evaluationssynthesen und gegebenenfalls auch Meta-Analysen möglich macht (DeGEval 2017).

3 Fazit

Standards sind nicht nur Hilfsmittel bei der Konzeption von Evaluationen, sie dienen auch als Grundlage für Bewertungen und sind zentral bei der Qualitätskontrolle. Evaluationsstandards können auch – durch eine angemessene Berücksichtigung in einzelnen Evaluationen – den Wissenstransfer im Themen- und Forschungsfeld der Evaluationen fördern. Es ist grundsätzlich möglich, dass einzelne Aspekte durch Evaluationsteams oder damit betrauten Personen – bspw. bei Selbstevaluationen – angepasst werden müssen oder sich nicht anwenden lassen (DeGEval 2004a). Da die Standards als sog. Maximalstandards formuliert sind, kann es in der Praxis bspw. zu Konflikten zwischen zwei oder mehr dieser Standards kommen, sodass begründete Einschränkungen erfolgen müssen. Jedoch sind Auslassungen und Anpassungen von Einzelstandards immer sorgfältig abzuwägen und schriftlich festzuhalten (DeGEval 2004a). Nützliche, durchführbare,

faire und genaue Evaluationen zu ermöglichen, ist jedoch nicht allein die Aufgabe von Evaluator:innen. „Für die Qualität von Evaluationen sind viele mitverantwortlich“ (Hense 2021, S. 5). Die fachlichen Standards für Evaluationen richten sich gleichermaßen an Auftraggebende, Entscheidungstragende, Projektbeteiligte, Zielgruppenangehörige und je nach Konstellation auch an weitere Akteur:innen (Hense 2021). Evaluationen können zudem Grundlagen für Fort- und Weiterbildungen sein und unterstützen eine nachvollziehbare Argumentation (DeGEval 2004b). Insbesondere im Handlungsfeld der Selbstevaluation hat die *DeGEval* weitere spezifische Konkretisierungen und Interpretationshilfen erarbeitet und empfiehlt die Anwendung einer besonderen Gewichtung einzelner Evaluationsstandards, die insbesondere die (Doppel-)Rolle der Evaluator:innen und die daraus abzuleitenden feldspezifischen Anforderungen betreffen (DeGEval 2004a).

Literatur

Beywl W (2019) Evaluationsstandards–Orientierungshilfen für Evaluationen in Schule und Unterricht. In: Buhren CG, Klein G, Müller S (Hrsg) Handbuch Evaluation in Schule und Unterricht. Beltz, Weinheim, S 30–44

DeGEval [Gesellschaft für Evaluation] (2004a) Empfehlungen zur Anwendung von Standards für Evaluation im Handlungsfeld der Selbstevaluation. DeGEval – Gesellschaft für Evaluation, Alfter

DeGEval [Gesellschaft für Evaluation] (2004b) Empfehlungen für die Aus- und Weiterbildung in der Evaluation. Anforderungsprofile an Evaluatorinnen und Evaluatoren. DeGEval – Gesellschaft für Evaluation, Mainz

DeGEval [Gesellschaft für Evaluation] (2017) Standards für Evaluation. Erste Revision 2016. DeGEval – Gesellschaft für Evaluation, Mainz

Hense JU (2021) Nicht jede Evaluation ist eine gute Evaluation. Warum gute Evaluationen fachliche Standards berücksichtigen sollten. FORUM Sexualaufklärung und Familienplanung: Informationsdienst der Bundeszentrale für gesundheitliche Aufklärung (BZgA) 1:3–5

Sanders JR (2013) Handbuch der Evaluationsstandards: Die Standards des Joint Committee on Standards for Educational Evaluation. VS Verlag, Wiesbaden

Westermann R (2016) Methoden psychologischer Forschung und Evaluation. Kohlhammer, Stuttgart

Vanessa van den Bogaert ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Lehr-Lernforschung der Ruhr-Universität Bochum. Sie widmet sich in ihren Forschungsschwerpunkten der wissenschaftlichen Begleitforschung von Citizen-Science-Projekten sowie der Grundlagenforschung zur Interessengenese an außerschulischen Lernorten. Sie leitet die Arbeitsgruppe *Science of Citizen Science* in Zusammenarbeit mit *Bürger schaffen Wissen*.


cc
BY

(Einzel-) Methoden der Evaluation von Wissenschaftskommunikation



Grundlagenbeitrag: Quantitative Befragungen

Christoph Böhmert und Ferdinand Abacioglu

Zusammenfassung

Die quantitative Befragung ermöglicht eine effiziente Erhebung und Auswertung großer Mengen an Daten und hat somit ihren festen Platz im Methodenkanon vieler empirischer Wissenschaften. Der vorliegende Beitrag setzt sich im Rahmen der Evaluation von Wissenschaftskommunikation mit den Grundlagen dieses Erhebungsverfahrens auseinander und geht dabei zunächst auf wesentliche Merkmale und Klassifizierungsmöglichkeiten der Methode ein. In Hinblick auf Konzeption und Durchführung werden dann zentrale Anforderungen an das entsprechende Messinstrument, sowie die grundlegenden Eigenschaften von Messskalen und Antwortformaten dargelegt. Der Beitrag gibt zudem praktische Hinweise für die Verwendung von Ratingskalen und schließt mit einer Zusammenfassung bewährter Praktiken bei der allgemeinen Fragenkonstruktion.

Häufig werden in der Forschung quantitative und qualitative methodische Ansätze voneinander abgegrenzt (Döring und Bortz 2016, S. 14 ff.). Wenngleich die Unterscheidung der beiden Ansätze eine lange Tradition hat, lassen sich die Ansätze auch in sogenannten Mixed-Methods-Ansätzen im Rahmen von Forschungsprojekten miteinander kombinieren und integrieren, sodass das

C. Böhmert (✉)
IU Internationale Hochschule, Karlsruhe, Deutschland
E-Mail: christoph.boehmert@iu.org

F. Abacioglu
IU Internationale Hochschule, Frankfurt am Main, Deutschland
E-Mail: ferdinand.abacioglu@iu.org

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_5

„jeweilige Forschungsproblem umfassender bearbeitet werden kann […] und eine bessere Absicherung der Ergebnisse möglich ist" (Döring und Bortz 2016, S. 17). Der vorliegende Beitrag beschäftigt sich mit quantitativen Befragungen, es sei aber explizit darauf verwiesen, dass auch im Rahmen von Befragungen quantitative und qualitative Methoden kombiniert werden können. Tatsächlich werden in der wissenschaftskommunikativen Evaluationspraxis sehr häufig quantitative Befragungen (etwa die Bewertung eines Vortrags, des Auftretens des Vortragenden etc. mit einer Schulnote) mit qualitativen Elementen (etwa die offene Frage danach, was am Vortrag besonders gut war und was hätte besser sein können) ergänzt.

Kernmerkmal quantitativer Sozialforschung (im Gegensatz zu qualitativer Sozialforschung) ist, dass numerische Daten, also Messwerte erhoben werden. Solche Messwerte können auf unterschiedliche Arten erhoben werden (Döring und Bortz 2016, S. 321 ff.): In der Evaluationsforschung von Bedeutung sind neben den hier thematisierten Befragungen auch Beobachtungen (z. B. wie sich die Zuschauer im Rahmen eines Science Slams verhalten, zu Beobachtungen allgemein siehe auch Weiß in diesem Band) und Tests (z. B. Wissenstests im Anschluss an auf Wissensvermittlung ausgelegte Formate der Wissenschaftskommunikation, siehe auch Wirth und Fleischer in diesem Band).

Bei quantitativen Befragungen wird in der Regel ein möglichst hoher Grad an *Standardisierung* der Befragung angestrebt (Reinecke 2019, S. 717). Diese bezieht sich auf den Grad der Festlegung

- des Fragetextes,
- der Antwortkategorien sowie
- der Reihenfolge der Fragen.

1 Taxonomie quantitativer Befragungen

Quantitative Befragungen können hinsichtlich verschiedener Aspekte klassifiziert werden. Diese Klassifikationen werden hier zunächst eingeführt und weiter unten in diesem Abschnitt an Beispielen aus der Evaluation von Wissenschaftskommunikation verdeutlicht.

Für die Durchführung einer Befragung lassen sich klassischerweise die drei Befragungsarten persönlich-mündlich, telefonisch und schriftlich ausmachen (Fuchs 2019). Neben diesen Arten einer Befragung (auch Modi genannt) sind allerdings insbesondere auch Onlinebefragungen und mobile Befragungen zu erwähnen. Zur weiteren Differenzierung von Befragungen werden in der Literatur

daher häufig die Dimensionen *Administrationsform, Kommunikationskanal* und *Befragungstechnologie* herangezogen (Faulbaum et al. 2009; Reinecke 2019). So identifiziert die Administrationsform einer Befragung, von wem die Fragen (vor-) gelesen beziehungsweise die Antworten registriert werden. Dies kann entweder durch die befragte Person selbst geschehen oder durch eine interviewende Person. Der primäre Kommunikationskanal kann visuell oder auditiv sein, wobei auch Mischformen möglich sind. Als dritte Dimension unterscheidet die Befragungstechnologie zunächst schlicht zwischen der reinen An- und Abwesenheit einer technologischen Unterstützung (Faulbaum et al. 2009). So kann beispielsweise eine telefonische Befragung ohne weitere Unterstützung durchgeführt werden oder durch den Einsatz entsprechender Software aufseiten der Interviewenden computergestützt sein. In Hinblick auf unterschiedliche Kommunikationstechnologien und Endgeräte – welche grundlegenden Einfluss auf die Art einer Befragung nehmen können – ist jedoch anzumerken, dass die Befragungstechnologie nicht als reine Unterstützung der bereits angesprochenen klassischen Modi verstanden werden soll: Je nach Befragungstechnologie lassen sich zusammen mit Administrationsform und Kommunikationskanal spezifische Befragungsarten abbilden.

Eine weitere Klassifizierungsmöglichkeit stellen die Befragten selbst dar: Wer wird überhaupt befragt? Bei der Evaluation von Wissenschaftskommunikation kann hier nach den Rezipient:innen, den Kommunikator:innen sowie weiteren Stakeholder:innen (z. B. Auftraggeber:innen) unterschieden werden. Innerhalb der Gruppe der Rezipient:innen kann nochmals zwischen Rezipient:innen, auf die die Kommunikation abzielt (Zielgruppe), und Rezipient:innen, auf die diese nicht explizit abzielt (sonstige Rezipient:innen), unterschieden werden. Der Rezipient:innenbegriff umfasst hier sowohl aktive, mit den Kommunikator:innen in Dialog stehende als auch passive Rezipient:innen. Abb. 1 visualisiert diese Taxonomie quantitativer Befragungen im Rahmen der Evaluation von Wissenschaftskommunikation.

Darüber hinaus können Befragungen auch hinsichtlich des Befragungssettings (Befragung von Einzelpersonen vs. einer Gruppe von Personen) sowie hinsichtlich des Befragungsgegenstands unterschieden werden (Raithel 2008).

Bei Befragungen in der Evaluation ebenfalls von großer Relevanz sind generelle Aspekte des Forschungsdesigns. Diese spielen auch bei anderen Evaluationsmethoden eine große Rolle. Zu nennen ist hier insbesondere die Planung des Untersuchungszeitpunkts bzw. auch mehrerer Untersuchungszeitpunkte (Döring und Bortz 2016).

Betrachten wir zur Veranschaulichung zwei Beispiele, in denen Wissenschaftskommunikation evaluiert wird:

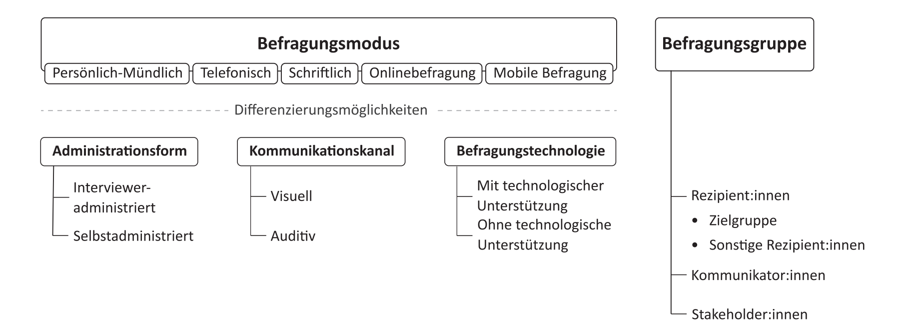


Abb. 1 Klassifizierungsmodell quantitativer Befragungen im Rahmen der Evaluation von Wissenschaftskommunikation. Befragungsmodus in Anlehnung an Faulbaum et al. (2009) und Reinecke (2019)

1. Bewertung von Science Slam-Beiträgen durch das Publikum
2. Bewertung der Verständlichkeit von wissenschaftsjournalistischen Texten

Beispiel 1: Am Ende eines Science Slams bewertet das Publikum alle gehaltenen Vorträge. Dazu findet sich das Publikum in mehreren Gruppen zusammen. In diesen Gruppen werden die Vorträge diskutiert und anschließend Punkte für den Vortrag (oder mehrere Punktwerte für unterschiedliche Aspekte des Vortrags) vergeben. Die jeweiligen Punkte werden den Organisator:innen der Veranstaltung anschließend durch das Hochhalten von Bewertungstafeln (mit den Nummern 1 bis 10) mitgeteilt. Wie lässt sich diese Form der Bewertung klassifizieren? Es handelt sich offenbar um eine standardisierte Befragung, da die Anweisungen, welche die Moderatorin über das Mikrofon gibt, für alle bewertenden Gruppen gleich sind, die Antwortkategorien (Tafeln mit Nummern von 1 bis 10) vorgegeben sind, und die Abfrage für sämtliche Gruppen in derselben Reihenfolge stattfindet (zunächst Bewertung von Vortrag 1, dann von Vortrag 2 etc.). Die Befragung ist durch eine interviewende Person (die Moderatorin) administriert. Der Kommunikationskanal bezieht sich nach Faulbaum et al. (2009) auf den Kanal, über den Befragte ihren Input erhalten. In unserem Science Slam Beispiel geschieht dieser Input im Wesentlichen über den auditiven Kanal (die Moderatorin formuliert die Frage mündlich), zum Teil aber auch über den visuellen (z. B. Mimik und Gestik der Moderatorin). Die Befragung geschieht persönlich und es wird die Gruppe der Rezipient:innen befragt.

Beispiel 2: In einer Studie von Böhmert et al. bewerteten Proband:innen (Studierende) wissenschaftsjournalistische Kurznachrichten unter anderem hin-

weitschweifig	+++	++	+	O	+	++	+++	aufs Wesentliche beschränkt

Abb. 2 Beispielitem gemäß der Verständlichkeitsskala von Langer et al. (2019)

sichtlich ihrer Verständlichkeit (Böhmert et al. 2021). Die Kurznachrichten sowie die Fragen dazu erhielten sie im Hörsaal am Ende einer Lehrveranstaltung. Zur Messung der Verständlichkeit wurde eine bestehende Skala von Langer et al. (2019) adaptiert. Die Skala besteht aus insgesamt 20 Items, die vier zugrunde liegende Dimensionen quantifizieren. Ein Beispiel-Item für die Dimension „Kürze/Prägnanz" ist in Abb. 2 gezeigt. Es handelt sich hier ebenfalls um eine standardisierte Befragung, bei der alle Teilnehmenden exakt dieselbe Instruktion, dieselben Fragen und Antwortmöglichkeiten sowie dieselbe Reihenfolge der Fragen erhielten. Die Instruktionen waren in Schriftform im Fragebogen gegeben, daher handelte es sich um eine selbstadministrierte Befragung über einen visuellen Kommunikationskanal. Die genutzte Befragungstechnologie war „persönlich" (ohne technische Hilfsmittel, einsammeln der Fragebogen am Ende der Vorlesung) und befragt wurden ebenfalls die Rezipient:innen.

2 Hauptgütekriterien

Um die interessierenden Merkmale eines Evaluationsgegenstandes zu untersuchen, bedarf es geeigneter Messinstrumente. Dabei kann die Qualität der Instrumente anhand mehrerer Gütekriterien überprüft werden: Im Rahmen der quantitativen Forschung werden darunter Anforderungen verstanden, welche sich im Wesentlichen auf Konzeption und Anwendung der Messinstrumente beziehen. Unterschieden werden die drei Hauptgütekriterien der Objektivität, Reliabilität und Validität. Diese Gütekriterien müssen im Rahmen von Befragungen soweit wie möglich sichergestellt werden. Nur dann kann davon ausgegangen werden, dass eine Befragung Daten von hoher Qualität liefert.

2.1 Objektivität

Ein grundlegender Anspruch an verwendete Messinstrumente ist die Vergleichbarkeit ihrer Ergebnisse: Werden Daten nicht immer auf die gleiche Weise

erhoben, ausgewertet und interpretiert, unterliegt die Messung Störeinflüssen, welche die gewonnen Ergebnisse verzerren können (Moosbrugger und Kelava 2020). Das Kriterium der Objektivität bezeichnet daher die Anforderung, dass äußere Einflüsse konstant gehalten werden müssen. Befragungen werden von unterschiedlichen Testleiter:innen immer gleich durchgeführt und ein gegebener Datensatz wird von ihnen immer auf die gleiche Weise ausgewertet und interpretiert. Hierfür können drei Dimensionen des Kriteriums unterschieden werden, welche sich auf diesen Prozess der Messung beziehen: Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität (Moosbrugger und Kelava 2020). Die objektive Durchführung einer Befragung erfordert in erster Linie eine Standardisierung der Fragetexte und der Reihenfolge der Fragen. Für eine objektive Auswertung ist unter anderem die Standardisierung der Antwortkategorien Voraussetzung. Ist auch die Interpretation der numerischen Ergebnisse hinsichtlich des zu evaluierenden Merkmals standardisiert, kann die Interpretationsobjektivität als gegeben angesehen werden.

2.2 Reliabilität

Im Gegensatz zur Objektivität, welche sich auf äußere Einflüsse bezieht, beschreibt Reliabilität eine Eigenschaft des Messinstruments beziehungsweise der Messung selbst. Von einem reliablen, d. h. zuverlässigen Instrument wird erwartet, dass es unter gleichen äußeren Bedingungen (gegebene Objektivität) für denselben, unveränderten Untersuchungsgegenstand immer wieder zu den gleichen Ergebnissen kommt – das bedeutet, dass Ergebnisse grundlegend reproduzierbar sind. Häufig ist eine exakte Wiederholung der Ergebnisse allerdings nicht möglich: In der Praxis unterliegen Messungen gemäß der sogenannten klassischen Testtheorie immer einem gewissen zufälligen Messfehler, welcher dafür sorgt, dass der gemessene Wert nicht exakt mit dem wahren Wert (beispielsweise exakte Meinung einer befragten Person) übereinstimmt (Moosbrugger und Kelava 2020). Je stärker ein Instrument messfehlerbehaftet ist, desto unzuverlässiger und unpräziser kann das Merkmal des Evaluationsgegenstandes nur untersucht werden und entsprechend gering ist die Reliabilität der Befragung. Doch wie kann die Reliabilität quantifiziert werden? Wird sie als Zuverlässigkeit eines Messinstrumentes im Sinne der Reproduzierbarkeit von Ergebnissen verstanden, kann eine Messung schlicht zu einem zweiten Zeitpunkt wiederholt werden. Je stärker die Ergebnisse beider Messungen miteinander zusammenhängen, desto geringer ist der Messfehler und entsprechend hoch die Reliabilität. Dieses als „Retest“ bezeichnete Verfahren beruht allerdings

auf der Annahme, dass der Untersuchungsgegenstand zwischen den zwei Messzeitpunkten unverändert bleibt. Einstellungen und Meinungen von Befragten, die im Zuge einer Evaluation erhoben werden, können sich jedoch über die Zeit verändern. In diesem Fall würde durch einen Retest nicht nur die Reliabilität, sondern auch der Grad besagter Einstellungs- oder Meinungsänderung erfasst. Neben der Retest-Reliabilität existieren mit der Paralleltestreliabilität, Split-Half-Reliabilität und Konsistenzanalyse weitere Verfahren zur Reliabilitätsbestimmung, welche keinem entsprechenden Zeiteffekt unterliegen. Eine Übersicht dieser klassischen Reliabilitätsmaße geben Moosbrugger und Kelava (2020); Cronbachs Alpha (Cronbach 1951) ist dabei eines der am häufigsten genutzten Maße. Der Koeffizient bildet die interne Konsistenz einer verwendeten Skala (eine Gruppe an Fragen, welche zusammen dasselbe Merkmal erfassen sollen) ab. Alpha gibt also darüber Auskunft, wie stark die einzelnen Fragen miteinander zusammenhängen. Wird angenommen, dass jede dieser Fragen einen eigenen „Test" darstellt und all diese Tests dasselbe Merkmal des Evaluationsgegenstandes erfassen, ergibt sich aus einer hohen internen Konsistenz auch eine hohe Reliabilität: Nur wenn der Messfehler der einzelnen als Test angenommenen Fragen gering ist, können sie auch hoch miteinander korrelieren. An dieser Stelle sei noch darauf hingewiesen, dass vor der Reliabilitätsbestimmung mit den oben beschriebenen Verfahren jeweils bestimmte messtheoretische Voraussetzungen überprüft werden müssen (siehe hierzu Gäde et al. 2020).

2.3 Validität

Als wichtigstes Gütekriterium hinsichtlich der praktischen Anwendung von Messinstrumenten kann die Validität (Gültigkeit) angesehen werden (Moosbrugger und Kelava 2020). Validität bedeutet, dass ein Test das misst, was er vorgibt zu messen (und nicht irgendetwas anderes). Objektivität und Reliabilität sind notwendige aber keine hinreichenden Voraussetzungen für die Validität (Moosbrugger und Kelava 2020). Eine quantitative Befragung kann also nur dann das messen, was sie messen soll (Validität), wenn sie unabhängig von äußeren Umständen (objektiv) und ausreichend zuverlässig (reliabel) Ergebnisse liefert. Es ist jedoch auch möglich, dass die Befragung, obwohl sie den Ansprüchen der Objektivität und Reliabilität genügt, in Wirklichkeit etwas anderes misst, als sie soll.

Evaluation von Wissenschaftskommunikation befasst sich mit der Qualität der Kommunikation. Doch was macht die Qualität von Wissenschaftskommunikation in einem bestimmten Kontext aus? Muss die Kommunikation

in erster Linie verständlich sein? Sollte sie zudem unterhaltsam sein? Sollte sie zum Handeln anregen? Diese und weitere Fragen müssen bei der Evaluation zunächst beantwortet werden, d. h. es muss zunächst geklärt werden, was eigentlich das Kommunikationsziel ist, und wie dieses gemessen werden soll. Erst im Anschluss daran können Evaluator:innen sich mit der Frage beschäftigen, wie sie eine quantitative Befragung so gestalten können, dass diese auch tatsächlich die Kommunikation evaluiert.

Bei der Sicherstellung der Validität sollten vier verschiedene Aspekte betrachtet werden (Moosbrugger und Kelava 2020). Ein zentraler Aspekt der Gültigkeit ist die sogenannte *Kriteriumsvalidität*. Diese ist dann gegeben, wenn ein Zusammenhang zwischen den Angaben einer Person in einer Befragung und ihrem Verhalten außerhalb der Befragungssituation besteht. In unserem Science-Slam-Beispiel wäre die Befragung dann kriteriumsvalide, wenn eine hohe Punktzahl des Vortrags mit anschließendem Verhalten einhergeht, wie etwa, dass die Befragten über gut bewertete Vorträge anschließend mehr recherchieren, diskutieren, diese weiterempfehlen etc. – und über schlecht bewertete Vorträge entsprechend nicht.

Ein zweiter, zentraler Aspekt ist die *Konstruktvalidität*. Konstruktvalidität ist dann gegeben, wenn die Struktur der Antworten auf den Fragebogen sowie deren Übereinstimmungen mit und Abgrenzung von Antworten auf anderen Messinstrumenten mit der zugrunde liegenden Theorie konform ist (Moosbrugger und Kelava 2020). Ein Aspekt im Rahmen der Evaluation von Wissenschaftskommunikation kann hier beispielsweise sein, dass man einen Zusammenhang zwischen der generellen Bewertung und der Bewertung der Verständlichkeit erwartet. Ein „perfekter" Zusammenhang wäre allerdings nicht zu erwarten, da man gemäß Theorie davon ausgehen würde, dass Verständlichkeit nur einen Teilaspekt der generellen Bewertung ausmacht.

Die beiden weiteren Aspekte der Validität, die *Augenscheinvalidität* und die *Inhaltsvalidität,* behandeln im Gegensatz zur Kriteriumsvalidität und Konstruktvalidität nicht die Antworten auf den Fragebogen, sondern den Fragebogen in seiner Gestaltung selbst. Augenscheinvalidität ist dann gegeben, wenn Laien davon ausgehen, dass ein Befragungsinstrument auch tatsächlich das misst, was es vorgibt zu messen. Inhaltsvalidität liegt dann vor, wenn auch Expert:innen zu diesem Schluss kommen.

3 Skalenniveau und Antwortformate

Auf Bitten der Moderatorin des Science Slams hält eine Gruppe von Zuschauer:innen ihre Punktzahl – in diesem Fall die Bestnote zehn – für einen Science Slam zum Thema Astrophysik in die Höhe. Technisch betrachtet findet in diesem Moment eine Messung statt, und zwar die Messung des Merkmals „Einschätzung der Vortragsqualität“. Merkmalsträger ist die Gruppe, die das Schild hochhält. Die gemessene Merkmalsausprägung ist in diesem Falle die zehn. Bei Messungen erfolgt ganz allgemein die Zuordnung eines numerischen Relativs (einer Zahl) zu einem empirischen Relativ (einer Merkmalsausprägung, z. B. „besonders positive Bewertung des Vortrags“; Wirtz und Nachtigall 2012). Vom zu messenden Merkmal selbst sowie von der Art der Messung hängt nun ab, welche Aussagen Evaluierende anhand ihrer Daten treffen können und welche Kennwerte (z. B. Mittelwert etc.) sie berechnen können. Man spricht in diesem Zusammenhang von verschiedenen Skalenniveaus. Das niedrigste Skalenniveau stellt die *Nominalskala* dar. Anhand nominalskalierter Messungen lassen sich lediglich Aussagen über Gleichheit oder Verschiedenheit treffen. Ein Beispiel ist die Messung des Merkmals Geschlecht mit dreistufiger Skala (0 = männlich, 1 = weiblich, 2 = anderes Geschlecht). Das nächsthöhere Skalenniveau ist die *Ordinalskala*. Bei ordinalskalierten Messungen spiegelt das numerische Relativ Verschiedenheit und zusätzlich eine Rangordnung im empirischen Relativ wider. Ein Beispiel hierfür wäre eine grobe Messung des Merkmals „Alter des Publikums" mit einer vierstufigen Skala (1 = unter 18, 2 = 18 bis 39, 3 = 40 bis 59, 4 = 60 und älter). Das über der Ordinalskala liegende Skalenniveau ist die *Intervallskala*. Intervallskalierte Messungen bilden im numerischen Relativ Verschiedenheit, Rangordnung und zusätzlich eine Gleichheit der Abstände (Äquidistanz) ab. Gehen wir in unserem Science Slam-Beispiel davon aus, dass der Unterschied im empirischen Relativ zwischen beispielsweise 2 Punkten und 4 Punkten gleich groß ist wie der zwischen 7 Punkten und 9 Punkten, dann gehen wir von einer intervallskalierten Messung aus (Hussy et al. 2013). Intervallskalierte Messung ist eine wichtige Voraussetzung vieler Auswertungsverfahren. Beispielsweise können Mittelwerte nur für mindestens intervallskalierte Merkmale[1] sinnvoll interpretiert werden. Das Skalenniveau ist ein entscheidendes Element, das bei der Entwicklung von quantitativen Befragungsinstrumenten mit bedacht werden muss.

[1] Das höchste Skalenniveau stellt die Verhältnisskala dar, die hier jedoch nicht weiter thematisiert werden soll (z. B. Rasch et al. 2009, S. 12).

Fragen werden mitunter anhand ihres Inhalts als auch ihrer Form unterschieden (Porst 2014). Inhaltlich können Fragen zu Einstellungen, Überzeugungen, Wissen, Verhalten oder Merkmalen einer befragten Person grob unterschieden werden. Die Form einer Frage bezieht sich hingegen darauf, wie eine Antwort darauf gegeben werden kann. Unterschieden werden hier geschlossene und offene Fragen, wobei auch Mischformen (halboffene Fragen) möglich sind, welche die zwei Fragetypen miteinander verbinden (Porst 2014). Geschlossene Fragen geben eine feste Anzahl an möglichen Antworten vor, aus welchen die befragte Person eine einzige (Einfachnennung) oder potenziell mehrere (Mehrfachnennung) auswählen kann. Ein Beispiel für eine geschlossene Frage mit Einfachnennung ist die Abfrage des Geschlechts. Dem gegenüber stehen offene Fragen, welche keine festen Antwortkategorien vorgeben. Im Rahmen quantitativer Forschung könnte hierfür die freie Angabe des Alters in Jahren als Beispiel dienen. Schlussendlich stellen halboffene Fragen eine Erweiterung des geschlossenen Typs dar, indem sie neben vorgegebenen Antwortalternativen zusätzlich die Option einer offenen Antwort ermöglichen. Geschlossene Fragen erlauben durch ihr vorgegebenes Antwortformat eine standardisierte Befragung. Neben der reinen Ein- und Mehrfachnennung von Antwortalternativen kann mit geschlossenen Fragen auch eine Abfrage von Einstellungen und Meinungen vorgenommen werden. In diesem Fall besitzen die Antwortkategorien eine Rangfolge: Die Wertungen von eins bis zehn, welche im Beispiel des Science Slams etwa für die allgemeine Qualität eines Vortrags vergeben werden, bilden einzelne Rangplätze ab und gelten daher als mindestens ordinalskaliert. Für die Auswertung solcher „Ratingskalen“ ist es günstig, wenn sie Intervallskalenniveau aufweisen. Abb. 3 zeigt eine Auswahl der vorgestellten Frageformen.

3.1 Skalenbeschriftung, Polarität und Anzahl der Antwortkategorien

Die erwähnte „Science-Slam-Skala“ von eins bis zehn Punkten beschränkt sich auf verbale Definitionen ihrer Endpunkte: Die Moderatorin erklärt, dass die eins für die schlechteste und die zehn für die beste Vortragsqualität steht. Soll in unserem Beispiel neben der allgemeinen Qualität des Vortrags auch der Unterhaltungswert erfasst werden, könnte ein Punkt das Urteil „sehr langweilig“ abbilden und zehn Punkte entsprechend das Urteil „sehr unterhaltsam“ – die Verbalisierung der Notenpunkte dazwischen bleibt allerdings aus. Eine Alternative zu dieser hauptsächlich numerischen Antwortskala bilden Antwortformate,

A Wie ist Ihr Geschlecht?
- ○ Männlich
- ○ Weiblich
- ○ Anderes Geschlecht

B Bitte geben Sie an, welche Bereiche der Forschung Sie am meisten interessieren.
Mehrfachnennung möglich.
- ☐ Life Sciences
- ☐ Biologie
- ☐ Chemie
- ☐ Medizin
- ☐ Anderes: ____________

C Der Inhalt des Science-Slam-Vortrags war unterhaltsam.

Trifft zu	Trifft eher zu	Trifft eher nicht zu	Trifft nicht zu
○	○	○	○

Abb. 3 Unterschiedliche Frageformen. **A** Geschlossenes Antwortformat mit Einfachnennung, **B** Halboffenes Antwortformat mit Mehrfachnennung und **C** Ratingfrage mit vierstufiger Antwortskala

für welche sämtliche Ränge explizit ausformuliert sind: So könnte im Rahmen des Science Slams die Aussage „Der Inhalt des Science-Slam-Vortrags war unterhaltsam" über die vierstufige Skala von „trifft zu" über „trifft eher zu" und „trifft eher nicht zu" bis hin zu „trifft nicht zu" bewertet werden. Eine entsprechende Verbalisierung ist für Befragte insgesamt leichter verständlich als rein numerische Skalen (Hussy et al. 2013). Zu berücksichtigen ist hier jedoch, dass die Vergabe solcher „Labels" bei allen Antwortkategorien nur für Skalen mit bis zu sieben oder teilweise auch neun Abstufungen sinnvoll ist (Franzen 2019).

In diesem Zusammenhang kann auch die Überlegung angestellt werden, wie viele Antwortkategorien überhaupt zur Verfügung stehen sollen. Wenige Abstufungen können zu einer verringerten Reliabilität führen, während zu viele Abstufungen keine weiteren Vorteile bringen und dazu tendieren, befragte Personen zu überfordern (Franzen 2019). In der Praxis haben sich daher vier- bis neunstufige Ratingskalen bewährt (Hussy et al. 2013). Während eine gerade Anzahl an Antwortkategorien die Tendenz zu einem der Skalenenden erzwingt („trifft eher zu" oder „trifft eher nicht zu"), besitzen ungerade Anzahlen eine neutrale Option in der Mitte – diese muss allerdings nicht zwangsläufig sinnvoll zu interpretieren sein: Befragte können damit angeben wollen, dass ihre Einstellung ambivalent ist („sowohl als auch") oder sie keine konkrete Einstellung

hinsichtlich der Frage besitzen („ich weiß nicht“) (Hussy et al. 2013). Um diese inhaltliche Ungenauigkeit aufzulösen, kann eine separate Antwortkategorie „ich weiß nicht“ bereitgestellt werden – allerdings zeigt sich, dass das Vorhandensein dieser Kategorie häufig zu einer verringerten Auseinandersetzung mit der Frage führt, und Befragte trotz eigener Meinung lieber auf verwertbare Angaben verzichten (Franzen 2019).

Schlussendlich ist zu klären, ob die Beschriftungen einer Skala bipolar, also mit zwei inhaltlich entgegengesetzten Endpunkten (z. B. „sehr langweilig“ bis „sehr unterhaltsam“), oder unipolar, also mit einem einzigen – jedoch unterschiedlich gewerteten – Endpunkt (z. B. „trifft zu“ bis „trifft nicht zu“), gewählt wird. Unipolare Skalen wie diese haben den Vorteil, dass ihre Dimensionen von Befragten nicht für jedes Item neu erfasst werden müssen, was die Befragung einfacher macht. In der Regel kommen Befragte damit besser zurecht (Franzen 2019).

4 Allgemeine Regeln zur Fragenkonstruktion

Bevor eine befragte Person verlässlich und wahrheitsgemäß antworten kann, muss von ihr zunächst sowohl die Semantik des Fragentextes als auch die Intention der Forschenden richtig interpretiert werden (Strack und Martin 1987). Wie gut Befragten dieser Prozess gelingt, bestimmt nicht zuletzt die Art und Weise, wie die gestellten Fragen formuliert wurden (Porst 2014). Die Qualität einer Frage hängt also unter anderem von der gewählten Formulierung ab – doch welche Aspekte sind hierfür zu berücksichtigen? Obwohl sich für die Fragenkonstruktion keine endgültigen Regeln aufstellen lassen, hat sich doch eine Art „Best Practice“ herauskristallisiert. Porst stellt in diesem Zusammenhang folgende zehn „Gebote“ auf (siehe Tab. 1).

Porsts Gebote beanspruchen allerdings keine Allgemeingültigkeit. Häufig kann es notwendig sein, gegen eines oder auch gegen mehrere Gebote zu verstoßen, wenn die konkrete Situation dies erfordert. So muss beispielsweise oft ein Kompromiss zwischen dem ersten, zweiten und zehnten Gebot gefunden werden, wenn Begriffe und Konzepte für das eindeutige Verständnis zunächst erklärt werden müssen. Die Gebote sind daher weniger als starre Regeln, sondern vielmehr als grundlegende „Wegweiser“ zu betrachten, welche für jede einzelne Frage erneut bedacht werden sollten (Porst 2014). So beschreibt Payne (1951) den Prozess der Fragenkonstruktion mehr als Kunst, denn als Wissenschaft – welcher erst durch individuelle Entscheidungen und Kreativität zu einer guten Frage führt.

Tab. 1 Zehn Gebote der Frageformulierung nach Porst (2014, S. 99 f.) – mit Erläuterungen

1. Du sollst einfache, unzweideutige Begriffe verwenden, die von allen Befragten in gleicher Weise verstanden werden!	… sonst sind Ergebnisse u. U. nicht vergleichbar
2. Du sollst lange und komplexe Fragen vermeiden!	… sonst werden Fragen schnell unverständlich und dafür anfällig, eines der restlichen Gebote zu brechen
3. Du sollst hypothetische Fragen vermeiden!	… sonst können Fragen nicht verlässlich beantwortet werden, da sie u. U. zu weit von der Lebensrealität Befragter entfernt sind
4. Du sollst doppelte Stimuli [d. h. zwei Befragungsgegenstände in einer Frage] und Verneinungen vermeiden!	… sonst können Fragen nicht wahrheitsgemäß beantwortet werden, da die Stimuli u. U. unterschiedlich beantwortet werden müssten oder Fragen werden bei insbesondere doppelter Verneinung schlicht missverstanden
5. Du sollst Unterstellungen und suggestive Fragen vermeiden!	… sonst können Fragen nicht wahrheitsgemäß beantwortet werden, da die Unterstellung u. U. nicht zutrifft oder Fragen können nicht verlässlich beantwortet werden, da u. U. eine Beeinflussung durch die Suggestion stattfindet
6. Du sollst Fragen vermeiden, die auf Informationen abzielen, über die viele Befragte mutmaßlich nicht verfügen!	… sonst können Fragen u. U. nicht wahrheitsgemäß beantwortet werden
7. Du sollst Fragen mit eindeutigem zeitlichen Bezug verwenden!	… sonst können Fragen nicht verlässlich beantwortet werden, da vage Zeitbezüge u. U. unterschiedlich verstanden werden
8. Du sollst Antwortkategorien verwenden, die erschöpfend und disjunkt (überschneidungsfrei) sind!	… sonst können Fragen nicht wahrheitsgemäß oder eindeutig beantwortet werden
9. Du sollst sicherstellen, dass der Kontext einer Frage sich nicht (unkontrolliert) auf deren Beantwortung auswirkt!	… sonst können Fragen nicht verlässlich beantwortet werden, da die vorangehende Frage oder Instruktion u. U. Einfluss auf die Beantwortung besitzt
10. Du sollst unklare Begriffe definieren!	… sonst können Fragen nicht verlässlich oder wahrheitsgemäß beantwortet werden

Abschließend sei darauf hingewiesen, dass die landläufige Meinung, Befragungen seien einfach zu konzipieren und umzusetzen, in Anbetracht der obigen Ausführungen mit Vorsicht zu genießen ist. Tatsächlich gibt es eine große Diskrepanz zwischen Fragen, die man im Alltag stellt, und soliden wissenschaftlichen Befragungen, wie sie im Rahmen wissenschaftskommunikativer Evaluationen zum Einsatz kommen sollten. Wissenschaftliche Befragungen erfordern somit eine gewissenhafte Planung und Vorbereitung – richtig umgesetzt ermöglichen sie im Rahmen der quantitativen Forschung allerdings auch die effiziente Erhebung beziehungsweise Auswertung großer Mengen an Informationen und schaffen die Vergleichbarkeit von Ergebnissen.

Literatur

Böhmert C, Niemann P, Hansen-Schirra S, Nitzke J (2021) Wen verstehen wir besser? Eine vergleichende Rezeptionsstudie zu Kurzmeldungen von Journalisten und Wissenschaftlern. In: Milde J, Welzenbach-Vogel IC, Dern M (Hrsg) Intention und Rezeption von Wissenschaftskommunikation. Halem, Köln, S 110–126

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334. https://doi.org/10.1007/BF02310555

Döring N, Bortz J (Hrsg) (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer, Berlin (Springer-Lehrbuch)

Faulbaum F, Prüfer O, Rexroth M (2009) Was ist eine gute Frage? VS Verlag, Wiesbaden

Franzen A (2019) Antwortskalen in standardisierten Befragungen. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien, Wiesbaden, S 843–854

Fuchs M (2019) Mode-Effekte. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien, Wiesbaden, S 735–744

Gäde JC, Karin S-E, Werner C (2020) Klassische Methoden der Reliabilitätsschätzung. In: Moosbrugger H, Kelava A (Hrsg) Testtheorie und Fragebogenkonstruktion. Springer, Berlin Heidelberg, S 305–334

Hussy W, Schreier M, Echterhoff G (Hrsg) (2013) Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor. Springer, Berlin (Springer-Lehrbuch)

Langer I, Schulz von Thun F, Tausch R (2019) Sich verständlich ausdrücken, 11. Aufl. Ernst Reinhardt Verlag, München.

Moosbrugger H, Kelava A (2020) Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien"). In: Moosbrugger H, Kelava A (Hrsg) Testtheorie und Fragebogenkonstruktion. Springer, Berlin, S 13–38

Payne SL (1951) The Art of Asking Questions. University Press, Princeton

Porst R (2014) Fragebogen. Springer Fachmedien, Wiesbaden

Raithel J (2008) Quantitative Forschung. Ein Praxiskurs, 2., durchgesehene Aufl. VS Verlag (Lehrbuch), Wiesbaden.

Rasch B, Friese M, Hofmann W, Naumann E (2009) Quantitative Methoden 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler, 3. Aufl. Springer, Heidelberg
Reinecke J (2019) Grundlagen der standardisierten Befragung. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien, Wiesbaden, S 717–734
Strack F, Martin LL (1987) Thinking, judging, and communicating: A process account of context effects in attitude surveys. In: Hippler HJ, Schwarz N, Sudman S (Hrsg) Social information processing and survey methodology. Springer, New York (Recent Research in Psychology), S 123–148
Wirtz MA, Nachtigall C (2012) Deskriptive Statistik. Statistische Methoden für Psychologen Teil 1. 6., überarb. Aufl. Beltz Juventa, Weinheim

Christoph Böhmert ist Professor für Kommunikationspsychologie an der IU Internationale Hochschule im Fernstudium. Im Feld der Wissenschaftskommunikation erforscht er schwerpunktmäßig die Kommunikation von Risiken so wie von Zahlen und Statistiken.

Ferdinand Abacioglu, M.Sc., ist Psychologe und derzeit als wissenschaftlicher Mitarbeiter an der IU Internationale Hochschule tätig. Dort übt er Lehrtätigkeiten im Studiengang Psychologie (B.Sc.) aus.





Praxisbeitrag: Quantitative Befragungen

Valerie Knapp und Vanessa van den Bogaert

Zusammenfassung

Für die Untersuchung partizipativer Formen der Wissenschaftskommunikation bzw. zur Erfassung von Eigenschaften der Teilnehmende dieser Maßnahmen können quantitative Befragungen einen wertvollen methodischen Zugang darstellen. Der Praxisbeitrag beleuchtet die praktische Umsetzung derartiger Befragungen am Beispiel mehrwelliger, onlinegestützter Panelstudien, die begleitend zu einer internationalen Citizen-Science-Aktion durchgeführt wurden. Die vorgestellten Befragungen zielten u. a. darauf ab, nähere Erkenntnisse darüber zu gewinnen, wer an der Aktion teilnahm und welche Erwartungen an eine Teilnahme geknüpft wurden. Zuletzt werden im Beitrag Vorteile und Limitationen des Ansatzes diskutiert, die Praktiker:innen in der Entscheidung zum Einsatz solcher Befragungen bzw. bei deren Planung und Umsetzung bedenken sollten.

V. Knapp (✉) · V. van den Bogaert
Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland
E-Mail: valerie.knapp@ruhr-uni-bochum.de

V. van den Bogaert
E-Mail: vanessa.vandenbogaert@ruhr-uni-bochum.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,*
https://doi.org/10.1007/978-3-658-39582-7_6

1 Einleitung: Quantitative Befragungen

Das Ziel der Wissenschaftskommunikation, einen reziproken Austausch zwischen Wissenschaft und Gesellschaft – jenseits der bloßen Vermittlung wissenschaftlicher Erkenntnisse – zu befördern, wird stetig wichtiger. Damit gewinnen auch partizipative Formen der Wissenschaft und der Wissenschaftskommunikation an Bedeutung. Diese Formate ermöglichen eine enge Kooperation von wissenschaftlichen und gesellschaftlichen Akteur:innen (Einsiedel 2008). Noch gibt es nur wenige generalisierbare Erkenntnisse beispielsweise darüber, wie eine erfolgreiche Zielgruppenansprache in und durch diese neuen Formate gelingen kann, was die Teilnahme an Selbigen motiviert, oder welche Erwartungen Personen mit einer Teilnahme verbinden. Dabei bieten sich verschiedene methodische Zugänge an, um mehr Erkenntnisse über die Teilnehmenden partizipativer Wissenschaftskommunikation zu gewinnen. Dieser Beitrag schärft aufbauend auf den Grundlagenbeitrag den Blick auf Methoden quantitativer Befragungen. Mit dem Ziel, Vor- und Nachteile dieser Form der Befragungen herauszuarbeiten, wird am Beispiel der Begleitforschung der europäischen Citizen-Science-Aktion *Plastic Pirates – Go Europe!* das Instrument mehrwelliger, onlinegestützter Panelstudien für die Evaluation von Maßnahmen der Wissenschaftskommunikation vorgestellt.

Der Beitrag bietet zunächst eine kurze Skizzierung des Konzepts von Citizen Science als Methode der Wissenschaftskommunikation. Daran anknüpfend liefert der Beitrag einen näheren Einblick in die Begleitforschung zu *Plastic Pirates – Go Europe!* sowie in die praktische Umsetzung einer Lehrer:innen-Befragung, die im Rahmen des Vorhabens umgesetzt wurden. Anhand dieses Beispiels werden abschließend die Vor- und Nachteile von Online-Befragungen zu zwei Messzeitpunkten für die Evaluation einer Maßnahme der Wissenschaftskommunikation reflektiert.

2 Citizen Science als Methode der Wissenschaftskommunikation

Externe Wissenschaftskommunikation – also die Kommunikation mit Adressat:innen, die außerhalb der Wissenschaft stehen wie z. B. Personen der breiteren Öffentlichkeit oder der Politik – hat in den vergangenen Jahrzehnten einen Wandel erlebt: Wissenschaftskommunikation beschreibt nicht länger die bloße Vermittlung wissenschaftlicher Erkenntnisse an die Öffentlichkeit, sondern zielt vielmehr auf den wechselseitigen Austausch zwischen wissenschaftlichen

und gesellschaftlichen Akteur:innen ab. Im Kontext dieses Wandels von einem unidirektionalen Wissenstransfer – welcher immer auch ein Informationsdefizit aufseiten der Öffentlichkeit unterstellt – hin zu Austausch und Kooperation auf Augenhöhe spielen partizipative Forschungsformate eine wichtige Rolle (Ball 2020). Zu diesen Formaten zählt auch die Bürgerforschung oder Citizen Science. Unter dem Begriff Citizen Science lassen sich eine Vielzahl verschiedener Forschungsformate fassen, die eine Einbindung von Bürger:innen in die Gewinnung wissenschaftlicher Erkenntnisse vorsehen (siehe auch Bruckermann und Greving sowie Greving et al. in diesem Band). Aus der Perspektive der Wissenschaftskommunikationsforschung geht die Beteiligung von Bürger:innen immer einher mit dem Ziel, die Wissensdifferenz zwischen Wissenschaftler:innen und Bürger:innen zu verringern (Bromme 2020). Die Beteiligung von Bürger:innen in realen Forschungskontexten schafft so einerseits die Möglichkeit, wissenschaftliches Arbeiten in und an der Praxis erfahrbar zu vermitteln. Andererseits bietet die Einbeziehung unterschiedlicher Personen und Gruppen in die Generierung, Interpretation und Diskussion wissenschaftlicher Erkenntnisse eine Lerngelegenheit für alle beteiligten Seiten gleichermaßen.

3 Praxisbeispiel: Onlinegestützte Lehrer:innen-Befragung im Rahmen der Begleitforschung *Plastic Pirates – Go Europe!*

Das Projekt Plastikpiraten, aus dem die Citizen-Science-Aktion *Plastic Pirates – Go Europe!* hervorgegangen ist, gibt es bereits seit 2016. Im Rahmen dieses Projektes sind Jugendliche der Jahrgangsstufe 9 dazu aufgerufen, im Klassenverband Plastikvorkommen in und an Fließgewässern nach wissenschaftlichen Vorgaben und innerhalb festgelegter Aktionszeiträume zu untersuchen. Dazu bestimmen und kartieren die Jugendlichen das Plastikvorkommen in verschiedenen Zonen in und an einem Fließgewässer. Die Aktionszeiträume liegen jeweils im Frühjahr und im Herbst. Ziel ist es herauszufinden, wie groß der Plastikeintrag in die Meere und Ozeane durch Flüsse ist. Im Zuge der gemeinsamen Trio-Präsidentschaft von Deutschland, Slowenien und Portugal im Rat der Europäischen Union in den Jahren 2020 und 2021 wurde die Aktion – gefördert durch das deutsche Bundesministerium für Bildung und Forschung (BMBF) – unter dem Titel *Plastic Pirates – Go Europe!* in allen drei Ländern durch Jugendliche umgesetzt (siehe auch http://plastic-pirates.eu). Die Jugendlichen wurden jeweils von ihren Lehrer:innen unterstützt und begleitet. Lehrer:innen nehmen in diesem Projekt die Rolle von Vermittler:innen ein, die

die standardisierte Durchführung der Aktion durch die Schüler:innen ermöglichen. Sie bereiten die Teilnahme mit den bereitgestellten Aktionsmaterialien so vor und nach, dass die Aktion in den entsprechenden Fachunterricht eingebettet werden kann. Im begleitenden Lehr- und Arbeitsmaterial finden Lehrkräfte Hintergrundinformationen zu Meeren und Ozeanen oder Kopiervorlagen für Aufgabenblätter. Zwar werden im Material bereits Hinweise zu Anknüpfungspunkten zum Kernlehrplan gegeben, die fachdidaktische Formulierung von bspw. Grob- und Feinlernzielen sowie die Anpassung an die vielfältigen Voraussetzungen der individuellen Schüler:innengruppe wird jedoch von den Lehrer:innen übernommen.

In der Literatur finden sich nur vereinzelt gesicherte Erkenntnisse über die Zielerreichung einzelner bürger:innenwissenschaftlicher Projekte im Bereich von Bildung und Vermittlung. Diese lassen darüber hinaus aufgrund methodischer Limitation derzeit kaum verallgemeinerbare Rückschlüsse zu. Für das Projekt *Plastic Pirates – Go Europe!* wurde daher eine Begleitforschung (ebenfalls gefördert durch das BMBF) angeregt. Neben der Untersuchung von Outcome-Variablen (Wirkungen bei den Zielgruppen) war die Begleitforschung auch darauf angelegt, nähere Erkenntnisse darüber zu gewinnen, wer an der Aktion teilnimmt und welche Erwartungen an Wissenschaft Lehrer:innen an eine Teilnahme knüpfen. Zwar sind es die Jugendlichen, die im Kontext der Aktion *Plastic Pirates – Go Europe!* eine konkrete Fragestellung bearbeiten und somit die zentrale Rolle bei der Datensammlung für die Citizen-Science-Aktion einnehmen, jedoch schaffen die Lehrer:innen den Rahmen, der eine Teilnahme der Jugendlichen erst ermöglicht. In der Begleitforschung wurde somit ein besonderes Augenmerk auf die Lehrer: innen in ihrer vermittelnden Rolle gelegt.

Ein erstes Ziel ist es, zu überprüfen, ob aus der Sicht der Lehrer:innen die Materialien eine authentische Vermittlung von wissenschaftlichen Arbeitsweisen ermöglichen. Die Erfassung von Authentizität in der Wissenschaftsvermittlung wird dazu auf Grundlage des theoretischen Modells zur Wahrnehmung von Authentizität in Lehr-Lernkontexten (Betz et al. 2016) mit den Authentizitätsdimensionen *Vermittler:in*[1], *Ort, Vorgehen* und *Innovation* erfolgen. Lehrer:innen wurden gebeten, einzuschätzen, für wie authentisch sie die durch die Schüler:innen im Rahmen der Aktionsteilnahme besuchten Orte sowie das

[1] Die Dimension Vermittler:in wird in dem Begleitforschungsvorhaben nicht untersucht, da die Lehrer:innen in diesem Projekt die Rolle der Vermittler:innen innehaben und somit eine Bewertung der Dimension nicht möglich ist.

eingesetzte Material (bspw. Proben, Geräte und Instrumente) halten und als wie wissenschaftlich sie die für die Aktion konzipierten Arbeits- und Vorgehensweisen wahrnehmen.

Ein zweites Ziel ist die empirische Untersuchung von Wissenschaftsvertrauen als Voraussetzung und als Ergebnis der Beteiligung in einem solchen Citizen-Science-Projekt. Aus theoretischer Sicht ist Vertrauen immer dort notwendig, wo ein arbeitsteiliges Aufeinander-angewiesen-Sein den Lernkontext kennzeichnet – wie dies im Kontext von Citizen-Science-Projekten auch der Fall ist. Allgemein kann Vertrauen entweder identifizierbaren Personen oder Institutionen, aber auch abstrakten Institutionen sowie Gruppen bzw. deren Stellvertreter:innen entgegengebracht werden (Bromme 2020). Wissenschaftsvertrauen wird im Rahmen der Begleitforschung sowohl als Voraussetzung für eine Teilnahme sowie als Ergebnis der Beteiligung bei der Citizen-Science-Aktion aufseiten der Lehrer:innen erfasst.

3.1 Vor der Datenerhebung – Planungsphase

Befragung zu mehreren Messzeitpunkten. Die Befragung der Lehrer:innen im Rahmen der Begleitforschung erfolgte über eine mehrsprachige Online-Erhebung im Design einer Panelstudie. Eine Panelstudie umfasst grundsätzlich die Erhebung bestimmter Merkmale individueller Beobachtungsträger:innen oder Befragungsteilnehmender zu unterschiedlichen bzw. mindestens zwei Zeitpunkten mittels standardisierter Fragebögen. Somit weist ein Panel Charakteristika sowohl der Erhebung von Querschnittsdaten als – über die Zeitreihe – auch von Längsschnittsdaten auf (Schröder 2007). Im Rahmen der Begleitforschung wurde mittels der Lehrer:innen-Befragung zu zwei Messzeitpunkten überprüft, ob eine Teilnahme an der Aktion *Plastic Pirates – Go Europe!* das Vertrauen der Lehrkräfte in Wissenschaftler:innen verändert.

Befragung in mehreren Ländern/Sprachen. Die Lehrer:innen-Befragung im Rahmen der Begleitforschung wurden auf Deutsch, Slowenisch und Portugiesisch umgesetzt, um länderübergreifende Gemeinsamkeiten sowie nationale Unterschiede bei der Implementation der Citizen-Science-Aktion nachzeichnen zu können. Zur Überprüfung auf Richtigkeit der Übersetzungen wird häufig ein Verfahren gewählt, in dem sich der Übersetzung eine Rückübersetzung anschließt, wobei die übersetzende Person den Originaltext nicht kennt (Bernard 2000). In diesem Praxisbeispiel überprüften die Partnerorganisationen in den beiden Ländern die durch einen Dienstleistenden angefertigten Übersetzungen hinsichtlich sprachlicher Korrektheit und Zielgruppenpassung.

Befragung anhand etablierter Instrumente. In der Lehrer:innen-Befragung wurden überwiegend etablierte Skalen genutzt (siehe Tab. 1). Das Verwenden etablierter Instrumente ist notwendig, wenn gesicherte Aussagen getroffen werden sollen, jedoch vorab Instrumente nicht selbst entwickelt und in separaten Evaluationsstudien geprüft werden können. Etablierte Erhebungsinstrumente können auf verschiedene Weise beschafft werden. Zum einen gibt es eine Vielzahl an Instrumenten, die online als Open-Access-Ressourcen verfügbar sind. Für solch etablierte Instrumente sind entsprechende psychometrische Eigenschaften wie Reliabilität und Validität in der Regel bereits empirisch überprüft und durch entsprechende Kennwerte dokumentiert, sodass eine Erhebung unter Verwendung solcher Instrumente die Gütekriterien guter Wissenschaftspraxis erfüllt.

Unabhängig davon sollten die psychometrischen Eigenschaften bei der Verwendung der Messinstrumente in einem neuen Kontext immer erneut überprüft und dokumentiert werden. Dazu wurden in diesem Praxisbeispiel die Fragen, auf denen die angestrebten quantitativen Aussagen beruhen, für die gesamte Stichprobe in gleicher Weise gestellt (Frageformulierung, Antwortformate, die Position der Frage im Fragebogen sowie Layout und Instruktionen waren identisch). Für die abhängigen Variablen der Lehrer:innen-Befragung wurden bereits publizierte bzw. im Publikationsprozess befindliche Fragebögen eingesetzt, die nach erfolgreicher Literaturrecherche und in den o. g. Quellen auch hinsichtlich der psychometrischen Eigenschaften als angemessen angesehen wurden. Bei den Variablen zur Stichprobenbeschreibung (demografische sowie professionsbezogene Angaben) wurden Items aus internationalen Large-Scale-Assessments (siehe auch Klingebiel und Klieme 2016) entlehnt, da sich diese speziell für den formellen Lernkontext, in der die Teilnahme der Aktion erfolgt, anbieten. Diese internationalen Studien können auch als möglicher Referenzrahmen zur Einordnung der professionsbezogenen Daten herangezogen werden. Insgesamt konnte ein hohes Maß an Standardisierung eingehalten werden (siehe dazu bspw. Schaeffer und Maynard 2008).

Pilotierung. Besonders in der Planungsphase sind alle wichtigen Entscheidungen zu treffen und divergierende Anforderungen abzuwägen. Eine Pilotierung der Studie ist notwendig, um bspw. herauszufinden, wie lange die geplante Befragung dauert, ob die Umsetzung technisch fehlerfrei erfolgt ist oder ob Teilnehmende die gestellten Fragen korrekt verstehen. Weiterhin gilt es, ein Optimum zwischen dem *workload* der Stichprobe und somit der Länge der Befragung und den zu beantwortenden Forschungs- bzw. Evaluationsfragen zu ermitteln. Als Faustregel gilt, dass Online-Befragungen nicht länger als 15 min dauern sollten (Bosnjak 2002). Die Anzahl der Fragen, die in dem vorgegebenen Zeitintervall bearbeitet werden können, wurden dazu vorab mit einer freiwilligen Stichprobe empirisch ermittelt.

Online-Befragungen. Die Durchführung der Erhebung als Online-Befragung ist hilfreich, wenn diese – wie auch im vorliegenden Praxisbeispiel – über einen längeren Zeitraum und über verschiedene Länder hinweg erfolgen soll. Hieraus ergeben sich jedoch Besonderheiten gegenüber einer analogen, schriftlichen Befragung einer Stichprobe, die bereits in der Pilotierungsphase mitgedacht werden müssen. Hier muss insbesondere ein sorgfältiges und langwieriges Testen der technischen Durchführbarkeit der Befragung auf unterschiedlichsten Endgeräten eingeplant werden. Ebenfalls empfehlenswert ist es, dass die Befragten über die gesamte Befragung hinweg erkennen können, wie weit sie im Fragebogen fortgeschritten sind, um so den erwarteten zeitlichen Anspruch besser einschätzen zu können (Schnell 2012). Im Praxisbeispiel der Begleitforschung wurde dies durch einen Fortschrittsbalken ermöglicht, denkbar wäre jedoch auch eine durchgehende Nummerierung der gestellten Fragen (z. B. „Frage 5 von 25"). Auch hinsichtlich Aufbau und Struktur einer Online-Befragung gilt es einiges zu beachten. Zunächst sollte die erste Seite – ähnlich dem Deckblatt einer analogen Befragung – alle wichtigen Informationen über die Befragung, wie beispielsweise deren Zweck und Dauer sowie Angaben über die durchführende Organisation bzw. Ansprechpartner:innen, deren Kontaktinformation sowie Erläuterungen zum Umgang mit den erhobenen Daten beinhalten (Schnell 2012). Offene Fragen sollten nach Möglichkeit am Ende der Befragung gestellt werden, da diese einen vorzeitigen Abbruch wahrscheinlicher machen (Schnell 2012). Nach Übermittlung der Antworten sollte den Teilnehmenden der Befragung für ihre Teilnahme gedankt werden.

Weiterhin muss das verwendete technische System den nötigen Anforderungen entsprechen und eine korrekte Durchführung der Online-Befragung auch bei zeitgleichem Zugriff durch eine große Zahl an Personen und bei Abruf auf unterschiedlichen Endgeräten funktionieren (Schnell 2012). Für die Lehrer:innen-Befragung im Rahmen der Begleitforschung wurde deshalb das kostenpflichtige[2], datenschutzkonforme Umfrage-Tool *Unipark* der *Tivian XI GmbH* genutzt.

Darüber hinaus muss technisch sichergestellt sein, dass jede befragte Person die Online-Umfrage nur einmal beantworten kann, jedoch ist es zugleich empfehlenswert, dass Unterbrechungen der Befragung möglich sind. Es ist darauf

[2] Neben kostenpflichtigen Umfrage-Tools gibt es auch kostenfreie Angebote oder solche, die auf ein gestaffeltes Preismodell setzen. Hier sind die (ökonomischen) Vorteile im Einzelfall gegen Nachteile z. B. hinsichtlich des möglichen Umfangs der Befragung abzuwägen.

zu achten, dass eine Wiederaufnahme nach einer solchen Unterbrechung an derselben Stelle in der Umfrage erfolgt (Schnell 2012). Im Falle der Lehrer:innen-Befragung wurden hierzu Cookies eingesetzt. Es ist wichtig zu beachten, dass die Verwendung von Cookies im Einklang mit der Europäische Datenschutz-Grundverordnung (DSGVO) hinweispflichtig ist.

Auch die Akquise der Stichprobe kann bei einer Online-Befragung auf unterschiedliche Weise erfolgen, abhängig davon, ob es sich um eine willkürliche Stichprobe oder eine echte Zufallsstichprobe handelt. Im Falle der Lehrer:innen-Befragung des hier vorgestellten Praxisbeispiels wurden die Einladungen zur Befragung mittels einer E-Mail-Liste[3] an interessierte Lehrer:innen versendet, wenn diese zuvor Aktionsmaterialien bestellt hatten. So konnten Lehrkräfte via eines per E-Mail an sie versendeten Links unmittelbar an der Befragung teilnehmen.

3.2 Durchführung

Die Lehrer:innen-Befragungen der Begleitforschung *Plastic Pirates – Go Europe!* wurden im Zeitraum von September 2020 bis Juni 2022 durchgeführt (siehe Abb. 1). Die Befragung der Stichproben fand zu jeweils zwei Messzeitpunkten statt, sodass die Befragten einmal vor und einmal nach Abschluss des Aktionszeitraums Angaben machten. Mittels eines persönlichen Codes, den die Teilnehmenden selbst generierten, konnten die pseudonymisierten Daten der verschiedenen Messzeitpunkte einzelnen Teilnehmenden zugeordnet werden. Die Grundgesamtheit für die Panelstudie bildeten alle Lehrer:innen, die für den entsprechenden Aktionszeitraum die erforderlichen Materialien auf der Aktions-Webseite bestellt hatten. Somit wurden zum einen Lehrer:innen befragt, die an der Aktion teilgenommen hatten. Zum anderen nahmen aber auch Personen an der Befragung teil, die ihr Interesse an der Citizen-Science-Aktion bekundet hatten, eine Teilnahme aus vielfältigen Gründen[4] aber nicht realisieren konnten.

---

[3] Die E-Mail-Liste wurde durch das *Ecologic Institut* verwaltet, welches das Organisationsbüro der Aktion *Plastic Pirates – Go Europe!* betreut. Die Begleitforschung hatte somit keinen unmittelbaren Zugriff auf die Adressen.

[4] Die Feldphase der Lehrer:innen-Befragung ging von Herbst 2020 bis Herbst 2021 und war daher in ihrer Umsetzung stark durch die COVID-19-Pandemie und die begleitenden Maßnahmen beeinflusst.

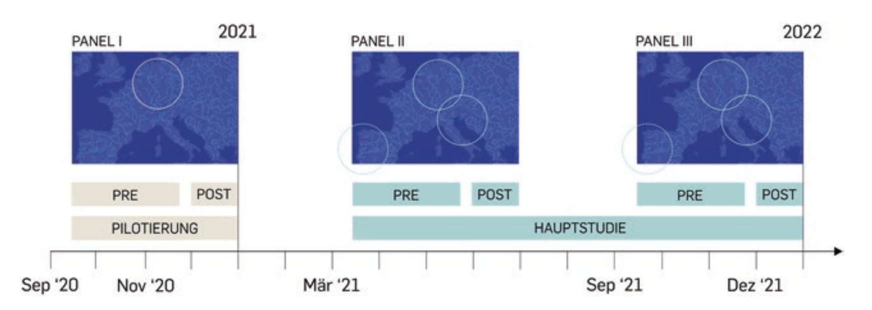


Abb. 1 Übersicht der Befragungszeiträume

3.3 Auswertung

Zugunsten von Hinweisen zur praktischen Umsetzbarkeit von Panelstudien wurde dem Bericht erster Ergebnisse nachfolgend etwas weniger Gewicht verliehen.[5] Der Aktionszeitraum Herbst 2020 (Panel I) ist als Pilotierung zu kennzeichnen. In dieser Phase wurden vor allem inhaltliche und technische Optimierungsbedarfe der begleitenden Aktion identifiziert.

Für die Lehrer:innen-Befragung gilt, dass die Teilnahme an der Aktion unabhängig von der freiwilligen Teilnahme an der Befragung erfolgen konnte. Zudem hatten einige Lehrkräfte bereits bei vorherigen Aktionszeiträumen (vor dem Herbst 2020) die Materialien bestellt und konnten somit nicht für die Befragung akquiriert werden (projektspezifische Besonderheit). Zwar wäre ein vollständiges Erreichen der Grundgesamtheit denkbar, jedoch ist eine solche Vollerhebung nicht immer möglich oder notwendig (siehe dazu auch Mayntz et al. 1969). Für die Datenerhebung im Rahmen der Pilotierung konnte eine zufriedenstellend große Stichprobe von $N = 119$ (prä und post) – in Relation zu den hochgeladenen Datensätzen des Aktionszeitraums von $N = 82$ – erreicht werden. Weiterhin zeigte die Pilotierung, dass die Befragung technisch fehlerfrei umgesetzt werden konnte und bspw. die geplante Befragungsdauer als angemessen einzuschätzen war.

Nach Hinweisen aus der Projektpraxis wurde die Strategie der Datenerhebung nach der Pilotierung für die Haupterhebung erweitert: Da nicht alle Lehrer:innen, die die Aktionsmaterialien herunterladen, eine Umsetzung der Citizen-Science-

[5] Weitere wissenschaftliche Publikationen in Fachzeitschriften über die Ergebnisse der Begleitforschung sind geplant.

Aktion tatsächlich realisieren und zudem die anhaltende COVID-19-Pandemie in einigen Fällen eine geplante Umsetzung unmöglich machte, wurde für die Lehrer:innen-Befragung eine Abzweigung für Nicht-Teilnehmende an der Citizen-Science-Aktion berücksichtigt, um auch deren Einschätzungen (wie z. B. die Gründe für eine Nicht-Teilnahme) zu erfassen. Für die Haupterhebungen (Panel II und III) wurden alle in dem jeweiligen Aktionszeitraum registrierten Lehrer:innen kontaktiert. Diese Datenerhebungsstrategie entspricht zwar der einer Panelstudie, jedoch nahmen nicht alle Befragten sowohl zum ersten als auch zum zweiten Messzeitpunkt an der Befragung teil. Die Erhebungen bedeuten aber für die explorative Nutzung weiterhin eine sinnvolle Datenbasis.[6] Während für die Überprüfung der psychometrischen Eigenschaften der eingesetzten Skalen (siehe Tab. 1) die Daten der gesamten Stichprobe aus Panel II und III aus dem Frühjahr und Herbst 2021 genutzt werden können, liegen für einen Vorher-Nachher-Vergleich nur Datensätze von Sub-Stichproben vor.

Für die Aktionszeiträume Frühjahr und Herbst 2021 (Panel II und III) wurden Daten für die deutsch- und slowenischsprachigen Stichproben erhoben.[7] Die Auswertung erfolgte über beide Zeiträume (Panel II und III) hinweg. Wie bereits beschrieben, liegen nicht für alle Teilnehmenden der Befragung die Daten für beide Messzeitpunkte vor. Zum ersten Messzeitpunkt der Hauptstudie ist ein Großteil der befragten Lehrer:innen in diesem Projekt weiblich und weist eine sehr heterogene Berufserfahrung in Jahren auf (Tab. 2).

3.4 Ergebnisse

Die Lehrer:innen bewerten das Vorgehen sowie die Arbeitsweisen als eine authentische Form der Wissenschaftsvermittlung (siehe Tab. 1). Dies bezieht sich einerseits auf die Einschätzung der zu bearbeitenden Fragestellung in dem Projektkontext[8] sowie dem wissenschaftlichen Vorgehen[9] und andererseits auf die Einschätzung der Authentizität des Ortes[10], an dem Schulklassen forschend tätig wurden.

---

[6] Dies gilt, sofern den unvollständigen Teilnahmen kein systematischer Befragungsabbruch (Dropout) zugrunde liegt.

[7] Eine Teilnahme an der Aktion konnte in Portugal aufgrund unterschiedlicher Faktoren wie u. a. der Corona-Pandemie nicht umgesetzt werden.

[8] Subskala Innovation $M\ (SD) = 4{,}34\ (0{,}61)$; Median = 4,3; Modus = 5.

[9] Subskala Vorgehen $M\ (SD) = 4{,}09\ (0{,}74)$; Median = 4; Modus = 4.

[10] Subskala Ort $M\ (SD) = 3{,}75\ (1{,}05)$; Median = 3,8; Modus = 5.

Tab. 1 Skalendokumentation METI und FEWAW

Konstrukt	Name des Instruments	Quelle	Kurzbeschreibung	Erhebungszeitpunkt	Ermitteltes Cronbachs Alpha	Mittelwert *(SD)*	Min – Max (Skala)
Vertrauen in Wissenschaftler:innen – prä	*Muenster Epistemic Trustworthiness Inventory* (METI)	Hendriks et al. (2015)	Eignet sich zur Erfassung von Einschätzungen von Vertrauenswürdigkeit. Mit 14 antonymen Adjektivpaaren (negativ – positiv) werden die Subdimensionen Expertise, Integrität und Wohlwollen gemessen. Semantisches Differential mit sieben Abstufungen.				
	METI Subskala Expertise	Hendriks et al. (2015)	6 Items	Hauptstudie prä	$n = 152$ $\alpha = 0{,}92$	6,0 (1,02)	1–7
				Hauptstudie post	$n = 124$ $\alpha = 0{,}92$	6,17 (0,96)	1–7
	METI Subskala Integrität	Hendriks et al. (2015)	4 Items	Hauptstudie prä	$n = 152$ $\alpha = 0{,}94$	5,5 (1,2)	1–7
				Hauptstudie post	$n = 124$ $\alpha = 0{,}92$	5,61 (1,10)	1–7
	METI Subskala Wohlwollen	Hendriks et al. (2015)	4 Items	Hauptstudie prä	$n = 152$ $\alpha = 0{,}93$	5,31 (1,16)	1–7
				Hauptstudie post	$n = 124$ $\alpha = 0{,}93$	5,47 (1,15)	1,5–7

(Fortsetzung)

Tab. 1 (Fortsetzung)

Konstrukt	Name des Instruments	Quelle	Kurzbeschreibung	Erhebungszeitpunkt	Ermitteltes Cronbachs Alpha	Mittelwert *(SD)*	Min – Max (Skala)
Authentizität – *post*	Fragebogen zur mehrdimensionalen Erfassung der Authentizitätswahrnehmung der Wissenschaftsvermittlung im Schülerlabor (FEWAW)	Finger et al. (2022)	Fragebogen zur Messung der Authentizitätswahrnehmung der Wissenschaftsvermittlung mit den Authentizitätsdimensionen Ort, Vorgehen und Innovation. 10 Items, fünfstufige Likertskala: „1 – stimme gar nicht zu" bis „5 – stimme voll und ganz zu"				
	FEWAW Subskala Ort	Finger et al. (2022)	3 Items	Hauptstudie post	$n=88$ $\alpha=0{,}89$	3,75 (1,05)	1–5
	FEWAW Subskala Vorgehen	Finger et al. (2022)	4 Items	Hauptstudie post	$n=88$ $\alpha=0{,}80$	4,09 (0,74)	2,3–5
	FEWAW Subskala Innovation	Finger et al. (2022)	3 Items	Hauptstudie post	$n=88$ $\alpha=0{,}80$	4,34 (0,61)	2,7–5

Tab. 2 Deskriptive Statistik

Erster Messzeitpunkt			
	N	Ja	Nein
Lehrkraft	152	91,4 %	8,6 %

	N	Männlich	Weiblich	Divers	Keine Angabe
Geschlecht	152	18,3 %	81,0 %	0 %	0,7 %

	N	M	SD	Min – Max
Alter in Jahren	152	44	9,7	26–72
Berufserfahrung in Jahren	139	14	9,3	0–50

Zweiter Messzeitpunkt			
	N	Ja	Nein
Lehrkraft	124	89,4 %	10,6 %

	N	Männlich	Weiblich	Divers	Keine Angabe
Geschlecht	124	21,8 %	77,4 %	0 %	0,8 %

	n	*M*	*SD*	Min–Max
Alter in Jahren	124	45,2	10,1	27–72
Berufserfahrung in Jahren	90	15,9	10,2	0–50

Für insgesamt $n = 45$ Befragte liegen Längsschnittdaten vor.[11] Anhand dieser Substichprobe wird das Vertrauen in Wissenschaftler:innen sowohl als Voraussetzung für eine Teilnahme sowie als Ergebnis der Beteiligung berichtet, um die Ausprägung vor und nach der Intervention innerhalb einer verbundenen Stichprobe vergleichend darstellen zu können. Lehrer:innen nehmen Wissenschaftler:innen mehrheitlich als kompetent, integer und wohlwollend

[11] Alter $M = 48{,}6$ Jahre ($SD = 10{,}6$); 77,8 % weiblich; 22 % männlich; 0 % divers; 2,2 % machten keine Angabe über ihr Geschlecht, mittlere Berufserfahrung in Jahren $M = 16{,}8$; $SD = 11{,}1$.

wahr. Insgesamt schätzen die befragten Lehrer:innen Wissenschaftler:innen somit als vertrauenswürdig ein.[12]

4 Vorteile von Befragungen in der Evaluation der Wissenschaftskommunikation

Befragungen mit mehreren Messzeitpunkten. Befragungen mit mehreren Messzeitpunkten bieten sich speziell bei der Untersuchung der Veränderung subjektiver Merkmale „wie z. B. Einstellungen, Kompetenzen und psychische Skalen“ (Pforr und Schröder 2015, S. 2) an, da diese zwar auch mittels Retrospektivfragen in Querschnittsstudien erhoben werden können, es hier jedoch z. B. durch verzerrte Erinnerungen der Befragten zu erheblichen Validitätsproblemen kommen kann (Pforr und Schröder 2015). Zudem können bei der Entscheidung zur Durchführung einer Panelstudie neben methodischen auch pragmatische Gründe eine Rolle spielen. So können sich Panelstudien beispielsweise anbieten, wenn im Rahmen einer Befragung eine große Anzahl von Fragen durch Individuen zu beantworten ist. Durch die Verteilung von Variablen, für die keine Veränderungshypothesen formuliert werden (wie bspw. demografische oder berufsbezogene Angaben), auf verschiedene Befragungswellen kann die Belastung der Befragten reduziert werden (Pforr und Schröder 2015). Im Falle der Lehrer:innen-Befragung im Rahmen der Begleitforschung wurden z. B. soziodemografische Angaben nur zum ersten Messzeitpunkt erhoben, da davon auszugehen war, dass diese über die Messzeitpunkte hinaus konstant blieben.

Online-Befragungen. Die Umsetzung einer Online-Befragung lässt gegenüber analogen, schriftlichen Erhebungen praktische Stärken erkennen. Zum einen ist eine Befragung schnell und ökonomisch umsetzbar, da die erhobenen Daten nach der Befragung nicht manuell eingegeben werden müssen, sondern unmittelbar nach der Erfassung digital zur Verfügung stehen (Schnell 2012). Die geringen Kosten für den Versand von E-Mails gegenüber anderen Modi in der Akquise sowie dem Nacherfassen der Stichprobe (Teilnahmeerinnerungen) stellen ins-

---

[12] Die drei – für die Subskalen separaten – Varianzanalysen mit Messwiederholung (A: Subskala Expertise $M\ (SD)_{prä} = 6{,}06\ (1{,}06)$; $M\ (SD)_{post} = 5{,}95\ (1{,}34)$; B: Subskala Integrität $M\ (SD)_{prä} = 5{,}71\ (1{,}14)$; $M\ (SD)_{post} = 5{,}49\ (1{,}30)$; C: Subskala Wohlwollen $M\ (SD)_{prä} = 5{,}49\ (1{,}13)$; $M\ (SD)_{post} = 5{,}38\ (1{,}31)$) mit Greenhouse–Geisser-Korrektur zeigen, dass es für das getroffene Vertrauensurteil keine statistisch signifikante Veränderung über das untersuchte Zeitintervall gibt, A: $F(1, 44) = 0{,}00$; B: $F(1, 44) = 1{,}30$, C: $F(1, 44) = 0{,}19$.

besondere bei der mehrmaligen Befragung, wie sie bspw. in Panelstudien erforderlich ist, einen merklichen Vorteil dar (Evans und Mathur 2005). Die einfache Überwindung geografischer sowie zeitlicher Barrieren durch die Durchführung einer Online-Befragung erlaubte es im Falle der Begleitforschung der Aktion *Plastic Pirates – Go Europe!*, die wiederholte Befragung bei relativ geringem personellem und finanziellem Aufwand in drei Sprachen zu adaptieren und somit kostengünstig international zu skalieren.

Ein Vorteil speziell von Online-Befragungen liegt in ihrer flexiblen Gestaltung und der Möglichkeit, verschiedenen Personen(-gruppen) nur die für sie relevanten Fragen anzuzeigen (Evans und Mathur 2005). Auch im hier vorgestellten Praxisbeispiel erhielten die Befragten auf sie zugeschnittene Umfragen, abhängig davon, ob sie z. B. Lehrer:innen waren oder nicht an der Aktion *Plastic Pirates – Go Europe!* teilgenommen hatten.

5 Nachteile von Befragungen in der Evaluation der Wissenschaftskommunikation

Befragungen mit mehreren Messzeitpunkten. Trotz vieler Vorteile haben Paneldesigns auch spezifische Nachteile, die es bereits bei der Planung der Erhebung zu bedenken gilt (van den Bogaert und Wirth 2020). Durch die freiwillige Teilnahme an den Panelbefragungen der Begleitforschung, die nicht an eine Teilnahme an der Aktion selbst geknüpft war, variierte die Stichprobe zu den einzelnen Messzeitpunkten, und zugleich lassen die erhobenen Daten der Stichproben nur beschränkt Rückschlüsse auf die Grundgesamtheit der an der Aktion beteiligten Lehrer:innen und Schüler:innen zu. Hier liegt auch schon eine der größten praktischen Schwierigkeiten einer Panelstudie: Die sogenannte Panelmortalität, die eine Nicht-Teilnahme zu allen Messzeitpunkten beschreibt. Dies wird vor allem dann zu einer methodischen Herausforderung, wenn die Ausfälle systematisch erfolgen und dadurch eine Verzerrung der Stichprobe auftritt (Diekmann 2009), da der Ausfall dann unmittelbar mit dem Untersuchungsgegenstand zusammenhängt (Tausendpfund 2018). Daher ist eine sorgfältige Kontaktpflege mit einer systematischen Kommunikationsstrategie wertvoll, um den Schwund im Panel durch fehlende oder sich ändernde Kontaktdaten möglichst zu minimieren. Insgesamt ist es ratsam, die erwartete Panelmortalität zuvor zu reflektieren und eine größere Stichprobe zu generieren (Schnell et al. 2018). Die Gründe für ein kleiner werdendes Panel sind vielfältig und nicht immer durch organisatorische Maßnahmen abwendbar. Jedoch sollten Art und Umfang der Befragung mit Blick auf Freude, Schwierigkeit oder Belastung für

die Teilnehmenden festgelegt werden, da diese Faktoren maßgeblich bei der Entscheidung sind, ob eine erneute Teilnahme erwogen wird (Hill und Willis 2001).

Neben der Panelmortalität können Erinnerungs-, Anpassungs-, Gewöhnungs- und Sättigungseffekte auftreten. Als sog. *Paneleffekte* (Schnell 2012) beschreiben sie das Phänomen, dass Teilnehmende aufgrund der wiederholten Befragung ihre Antworten ändern. Paneleffekte sind vor allem bei mehreren relativ umfangreichen Wiederholungserhebungen in kurzen Zeitabständen zu erwarten (Lohaus und Vierhaus 2015).

Online-Befragungen. Auch aus der Gestaltung der Befragung als Online-Befragung ergeben sich einige spezifische Nachteile. Beispielsweise birgt die Wahl einer selbstrekrutierten Stichprobe, wie sie auch im Falle der Panelstudien im Rahmen der Begleitforschung gezogen wurde, die Schwierigkeit, dass nur bedingt Schlüsse über die Grundgesamtheit gezogen werden können. Dies liegt vorrangig daran, dass die Befragten sich freiwillig zur Teilnahme an der Umfrage entschließen und damit keine Aussagen darüber möglich sind, ob und inwieweit sich diese von jenen Personen unterscheiden, die sich nicht zur Teilnahme an der Befragung entschlossen haben (Schnell 2012).

Zuletzt kann die Durchführung einer Panelstudie als Online-Befragung die beschriebene Schwierigkeit der Panelmortalität verstärken. Nicht nur der vorzeitige Abbruch, sondern auch das Ausscheiden zwischen den Messzeitpunkten sind bei einer Online-Befragung zumindest gegenüber einer persönlichen Befragung wahrscheinlicher. Im vorgestellten Praxisbeispiel wurde versucht, dieser Herausforderung durch die intensive Betreuung der befragten Lehrer:innen per E-Mail oder per Telefon zu begegnen.

6 Fazit

Mit der Zunahme partizipativer Formen der Wissenschaftskommunikation und deren wachsender Bedeutung gilt es für Evaluationsvorhaben, gesicherte Erkenntnisse über die Teilnehmenden dieser Formate sowie beispielsweise über deren Einstellungen und Erwartungen zu generieren. Quantitative Erhebungsmethoden können einen wertvollen Zugang zu derartigen Fragestellungen bieten. Für die Citizen-Science-Aktion *Plastic Pirates – Go Europe!* konnte mittels der hier vorgestellten Panelbefragung der Lehrkräfte die Fragen danach, wer an einer solchen Kampagne teilnimmt, systematisch näher beantwortet werden, wobei die Ergebnisse in diesem Beitrag nur ausschnittsweise berichtet werden konnten.

Neben der hier vorrangig fokussierten Nutzung der erhobenen Daten zur Beschreibung von Eigenschaften der Teilnehmenden, können Paneldaten durch

die Erhebung bestimmter Merkmale zu mindestens zwei Messzeitpunkten auch zur Überprüfung von Hypothesen dienen. Dennoch sollten die Herausforderungen einer Panelbefragung trotz ihrer methodischen und praktischen Stärken bei der Planung eines solchen Forschungsvorhabens genau reflektiert werden.

Danksagung Die Begleitforschung zur Citizen-Science-Aktion *Plastic Pirates – Go Europe!* wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01DRP2015 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen.

Literatur

Ball R (2020) Wissenschaftskommunikation im Wandel. Von Gutenberg bis Open Science. Springer VS, Wiesbaden

Bernard R (2000) Social research methods: qualitative and quantitative approaches. Sage, Thousand Oaks

Betz A, Flake S, Mierwald M, Vanderbeke M (2016) Modelling authenticity in teaching and learning contexts a contribution to theory development and empirical investigation of the construct. In: Looi CK, Polman J, Cress U, Reimann P (Hrsg) Transforming learning, empowering earners: The International Conference of the Learning Sciences (ICLS), Bd 2. Singapur, International Society of the Learning Sciences, S 815–818

Bosnjak M (2002) (Non)Response bei Web-Befragungen. Shaker, Aachen

Bromme R (2020) Informiertes Vertrauen: Eine psychologische Perspektive auf Vertrauen in Wissenschaft. In: Jungert M, Frewer A, Mayr E (Hrsg), Wissenschaftsreflexion. Interdisziplinäre Perspektiven zwischen Philosophie und Praxis. Mentis Verlag, Paderborn, S 105–134

Diekmann A (2009) Empirische Sozialforschung. Rowohlt, Reinbeck bei Hamburg

Einsiedel EF (2008) Public participation and dialogue. In: Bucchi M, Trench B (Hrsg) Handbook of public communication of science and technology, Routledge, London, S 173– 184

Evans JR, Mathur A (2005) The value of online surveys. Internet Res 15(2):195–219

Finger L, van den Bogaert V, Sommer K, Wirth J (2022) Das Schülerlabor als Ort authentischer Wissenschaftsvermittlung? Entwicklung und Validierung eines Fragebogens zur Erfassung der Authentizitätswahrnehmung der Wissenschaftsvermittlung im Schülerlabor. Zeitschrift für Didaktik der Naturwissenschaften, 28(2). https://doi.org/10.1007/s40573-022-00139-4

Hendriks F, Kienhues D, Bromme R (2015) Measuring laypeople's trust in experts in a digital age: the Muenster Epistemic Trustworthiness Inventory (METI). PLoS ONE 10(10):e0139309. https://doi.org/10.1371/journal.pone.0139309

Hill DH, Willis RJ (2001) Reducing panel attrition: a search for effective policy instruments. J Hum Resour 36(3):416–438

Klingebiel F, Klieme E (2016) Teacher qualifications and professional knowledge. In: Kuger S, Klieme E, Jude N, Kaplan D (Hrsg), Assessing contexts of learning. Springer, Cham, S 447– 468

Lohaus A, Vierhaus M (2015) Entwicklungspsychologie des Kindes- und Jugendalters für Bachelor. Springer, Berlin. https://doi.org/10.1007/978-3-662-45529-6_3

Mayntz R, Holm K, Hübner P (1969) Methoden der Stichprobenkonstruktion. In: Mayntz R, Holm K, Hübner P (Hrsg), Einführung in die Methoden der empirischen Soziologie. VS Verlag, Wiesbaden. https://doi.org/10.1007/978-3-322-96383-3_3

Pforr K, Schröder J (2015) Warum Panelstudien. Mannheim, GESIS Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://doi.org/10.15465/gesis-sg_008

Schaeffer NC, Maynard DW (2008) The contemporary standardized survey interview for social research. In: Conrad FG, Schober MF (Hrsg) Envisioning the survey interview of the future. Wiley, Hoboken, S 31–57

Schnell R (2012) Survey-Interviews: Methoden standardisierter Befragungen. VS Verlag, Wiesbaden

Schnell R, Hill PB, Esser E (2018) Methoden der empirischen Sozialforschung, 11., überarbeitete Aufl. De Gruyter Oldenbourg, Berlin, XIV, S 534

Schröder A (2007) Prinzipien der Panelanalyse. In: Albers S, Klapper D, Konradt U, Walter A, Wolf J (Hrsg) Methodik der empirischen Forschung. Gabler. https://doi.org/10.1007/978-3-8349-9121-8_18

Tausendpfund M (2018) Quantitative Befragungen in der Politikwissenschaft. Eine Einführung. Springer, Berlin

van den Bogaert V, Wirth J (2020) Praxisbeitrag – Panelbefragung als Instrument der Veränderungsmessung am Beispiel der Interessenentwicklung. In: Sommer K, Wirth J, Vanderbeke M (Hrsg) Handbuch Forschen im Schülerlabor – Theoretische Grundlagen, empirische Forschungsmethoden und aktuelle Anwendungsgebiete. Waxmann, Münster, S 257–266

Valerie Knapp ist wissenschaftliche Mitarbeiterin an der Ruhr-Universität Bochum. Ihre Forschungsschwerpunkte liegen u. a. in der Erforschung offener Wissenschaft sowie in der Begleitforschung und der Wirkungsevaluation von Citizen Science.

Vanessa van den Bogaert ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Lehr-Lernforschung der Ruhr-Universität Bochum. Sie widmet sich in ihren Forschungsschwerpunkten der wissenschaftlichen Begleitforschung von Citizen-Science-Projekten sowie der Grundlagenforschung zur Interessengenese an außerschulischen Lernorten. Sie leitet die Arbeitsgruppe *Science of Citizen Science* in Zusammenarbeit mit *Bürger schaffen Wissen*.


cc
BY



Grundlagenbeitrag: Qualitative Befragungen im Kontext von Wissenschaftskommunikation

Julia Metag und Andreas M. Scheu

Zusammenfassung

Der Beitrag beleuchtet das Potenzial qualitativer Befragungen zur Evaluation von Wissenschaftskommunikation. Qualitative Befragungen bieten sich immer dann an, wenn möglichst offen die Perspektiven, Bewertungen und Einschätzungen bestimmter Zielgruppen erhoben werden sollen. Insbesondere bei der Entwicklung neuer Formate und Formen von Wissenschaftskommunikation bieten sich narrative Befragungsformate zur Evaluation an. In der Forschung zeigt sich, dass halbstandardisierte Leitfadeninterviews sowie Fokusgruppen-Interviews sehr verbreitet sind. Qualitative Befragungen werden eingesetzt, um Expert:innen bzw. Wissenschaftler:innen, Bürger:innen und Vertreter:innen der Zivilgesellschaft sowie praktische Wissenschaftskommunikator:innen und Wissenschaftsjournalist:innen zu befragen. Qualitative Befragungen werden sowohl als primäre Forschungsmethode als auch in Kombination mit bzw. als Ergänzung zu standardisierten Methoden eingesetzt.

J. Metag (✉)
Institut für Kommunikationswissenschaft, Westfälische Wilhelms-Universität Münster, Münster, Deutschland
E-Mail: julia.metag@uni-muenster.de

A. Scheu
Berlin-Brandenburgische Akademie der Wissenschaften, Transfer Unit Wissenschaftskommunikation, Berlin, Deutschland
E-Mail: andreas.scheu@bbaw.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_7

1 Methode der qualitativen Befragung im Kontext der Evaluationsforschung

Die Stärken von qualitativen Befragungen liegen vor allem darin, Sinnzusammenhänge, Einschätzungen, Bewertungen oder Motive von Befragten zu erheben und zu verstehen. Damit bieten qualitative Befragungen auch große Potenziale für Evaluationen, zum Beispiel bei der Entwicklung neuer oder zur Weiterentwicklung bereits eingeführter Wissenschaftskommunikationsformate (siehe auch Hedder et al. in diesem Band). Grundsätzlich kommen dabei unterschiedliche Arten qualitativer Befragungen zum Einsatz, zum Beispiel offene Fragen in schriftlichen Fragebögen, mündliche Interviews (Helfferich 2014; Loosen 2016) bzw. Expert:inneninterviews (Blöbaum et al. 2016), Gruppendiskussionen (Lüthje 2016; Vogl 2014) oder auch DELPHI-Befragungen[1] (Häder und Häder 2014), die sich an unterschiedliche Zielgruppen richten können, zum Beispiel an Expert:innen, Kommunikator:innen oder Rezipierende im Bereich Wissenschaftskommunikation. Qualitative Befragungen werden sowohl als primäre Forschungsmethode als auch im Sinne einer Triangulation von Methoden in Kombination mit bzw. als Ergänzung zu standardisierten Methoden wie quantitativen Befragungen eingesetzt, um zum Beispiel Erkenntnisse aus repräsentativen Umfragen weiter vertiefen zu können (Koch et al. 2020; Koso 2021).

Dabei unterscheiden sich qualitative Befragungen im Grad der Standardisierung. Das Spektrum reicht hierbei von stärker strukturierten, teilstandardisierten Befragungsformen bis hin zu offenen, narrativen Formaten (Scholl 2009). Teilstandardisierte mündliche Befragungsformen nutzen mehr oder weniger ausführliche Leitfäden, die Schwerpunkte, Themen, den groben Gesprächsverlauf sowie Fragen und zum Teil auch Rückfragen strukturieren. So wird sowohl die Vergleichbarkeit der Interviews erhöht als auch sichergestellt, dass unterschiedliche Interviewende oder Moderator:innen von Gruppendiskussionen dieselben Inhalte ansprechen. Offene Befragungen, bei denen zum Beispiel nur eine Einstiegsfrage vorgegeben oder zu Beginn ein Stimulus gesetzt

[1] Hierbei handelt es sich meist um einen Mix aus qualitativen und quantitativen Befragungen, die wiederholt durchgeführt werden. Befragt werden Gruppen von Expert:innen zu Themen, Problemen und Sachverhalten, über die unsicheres und unvollständiges Wissen vorliegt. DELPHI-Befragungen werden beispielsweise eingesetzt, um Ideen zu generieren, Expert:innenwissen zu bündeln und zu bewerten oder auch um Prognosen zu erstellen.

wird, haben den Vorteil, dass die Befragten selbst bestimmen, welche Themen im Gespräch vertieft werden (Küsters 2014). Individuelle Schwerpunktsetzungen können dann genauso wie die jeweilige Strukturierung des Gesprächs von den Forschenden ausgewertet werden. So sind beispielsweise Rückschlüsse auf Relevanzzuschreibungen und Deutungen der Befragten möglich.

Je nach Zielsetzung der jeweiligen Evaluationsstudie bieten sowohl teilstandardisierte leitfadengestützte als auch narrative Befragungen besondere Stärken und Schwächen, die bei der Planung der Evaluation abgewogen werden sollten. Je wichtiger die Vergleichbarkeit sowohl zwischen den durchgeführten Befragungen und gegebenenfalls auch zwischen unterschiedlichen Untersuchungszeitpunkten ist, desto sinnvoller erscheint es, den Standardisierungsgrad zu erhöhen. Das ist beispielsweise dann der Fall, wenn bestimmte Kommunikationsformate bereits eingeführt wurden und im Zeitverlauf optimiert werden sollen. Wenn Kommunikationsformate neu entwickelt werden, können offene Befragungen wertvolle Anregungen und Perspektiven eröffnen. Hierbei spielt auch das verfügbare Vorwissen eine Rolle. Grundsätzlich bieten sich deduktive Vorgehensweisen – z. B. Konstruktion von analytischen Kategorien und Interviewleitfäden auf Basis des Forschungsstandes (Löblich 2016) – dann an, wenn bereits Erkenntnisse vorliegen und vertieft bzw. erweitert werden sollen. Je geringer das verfügbare Vorwissen ist, desto naheliegender ist der Einsatz offener, induktiver Zugänge zum Gegenstand, zum Beispiel in der Tradition der Grounded Theory (Scheu 2016).

Unabhängig davon, welche Art der qualitativen Befragung bei der Evaluation von Wissenschaftskommunikation eingesetzt wird, müssen spezifische, der Art der Befragung angemessene Qualitätskriterien beachtet und eingehalten werden. In der qualitativen Forschung sind die klassischen Gütekriterien empirischer Sozialforschung (Reliabilität, Validität und Repräsentativität; vgl. Wirth und Fleischer in diesem Band) zum Teil umstritten (Flick 2007, S. 499 f.). Uns erscheint die Bezugnahme auf diese Kriterien allerdings durchaus sinnvoll, allein schon, um die Aussagekraft der Evaluation und die davon abzuleitenden Konsequenzen zu legitimieren. Dabei müssen die Kriterien allerdings an die Logiken qualitativer Forschungsdesigns angepasst werden. Evaluationen, die qualitative Befragungen nutzen, wollen so den evaluierten Gegenstand, zum Beispiel ein Wissenschaftskommunikationsformat, weiterentwickeln. Dazu sind reliable, valide und dem Gegenstandsbereich angemessene Ergebnisse notwendig. Die Prüfung von Reliabilität und Validität steht in der qualitativen Forschung aber nicht am Ende des Forschungsprozesses. Zentrale Verfahrensweisen der qualitativen Befragung sollen helfen, reliable und valide Ergebnisse zu produzieren. Dazu gehört beispielsweise die kommunikative

Validierung von Interpretationsschritten in Teams, oder im Gespräch mit den Interviewpartner:innen, die nachvollziehbare und transparente Dokumentation von Entscheidungen und Schlussfolgerungen sowie die Reflexion der eigenen Schlussfolgerungen während des gesamten Forschungs- bzw. Evaluationsprozesses. Repräsentativität in Bezug auf qualitative Befragungen bezieht sich auf die bewusste Auswahl von Interviewpartner:innen (Theoretical Sampling). „Konzeptionelle Repräsentativität“ (Strübing 2008, S. 82) entsteht auch bei Evaluationen aus der wechselseitigen Bezugnahme von Schlussfolgerungen auf die erhobenen Daten und umgekehrt. Schlussfolgerungen von Evaluationen, die sich auf qualitative Befragungen berufen, haben also keinen probabilistischen Geltungsanspruch.

Die Stärken und der Mehrwert qualitativer Befragungen im Kontext von Evaluationen liegen stattdessen gerade in der Offenheit und damit verbunden der Möglichkeit, neue Perspektiven zu entdecken – dazu zählen auch unerwartete Rückmeldungen zu Aspekten, die in der Anlage der Evaluation eventuell überhaupt nicht berücksichtigt waren. Qualitative Befragungen können in Evaluationen eingesetzt werden, um die Sichtweisen, Deutungen und Bewertungen unterschiedlicher Zielgruppen zu rekonstruieren und zu verstehen und so einen besseren Zugang zum Objekt der Evaluation selbst zu bekommen.

2 Qualitative Befragungen im Kontext von Wissenschaftskommunikation

In der Wissenschaftskommunikationsforschung generell, nicht nur zur Evaluation von Wissenschaftskommunikationsmaßnahmen oder -formaten, werden qualitative Befragungen eingesetzt, um verschiedene Akteur:innen der Wissenschaftskommunikation und deren Einstellungen zu untersuchen. Das Sampling der Befragten funktioniert meist über ein konkretes Projekt oder einen konkreten Anlass (z. B. Almeida und Vaz Bevilaqua 2021) bzw. ein spezifisches wissenschaftliches Thema (Sharon und Baram-Tsabari 2020) oder auch über Listen von entsprechenden Wissenschaftsorganisationen oder Organisationen der praktischen Wissenschaftskommunikation. Es handelt sich vielfach also um gezielte Auswahlverfahren (Hassan et al. 2020; Samuel et al. 2021). Diese Sampling-Strategien werden häufig durch das Schneeballsystem ergänzt (Dudo et al. 2021).

In den meisten Fällen wird das halbstandardisierte Interview als Methode eingesetzt, d. h. es gibt einen vorbereiteten Interviewleitfaden, der allerdings flexibel an die Befragungssituation angepasst werden kann (z. B. Almeida und

Vaz Bevilaqua 2021; Dudo et al. 2021). Die Anzahl der Befragten ist unterschiedlich, sie liegt häufig zwischen etwa 10 und 30 Personen (z. B. Almeida und Vaz Bevilaqua 2021; Dudo et al. 2021; Llorente et al. 2021; Mahl et al. 2020; Sharon und Baram-Tsabari 2020), selten gibt es Studien wie die von Jones et al. (2020), in der 123 Personen qualitativ interviewt werden. Ein Interview dauert normalerweise zwischen 30 min und bis zu zwei Stunden (Almeida und Vaz Bevilaqua 2021; Dudo et al. 2021; Koso 2021; Sarathchandra und Haltinner 2020).

Darüber hinaus werden in der Wissenschaftskommunikationsforschung Fokusgruppen-Diskussionen eingesetzt, um durch die Gruppendynamik zu verstehen, wie sich Wissen und Ideen zu einem bestimmten Gegenstand oder Thema entwickeln, was auch gerade bei wissenschaftlichen Themen relevant sein kann (Asplund 2020; Hassan et al. 2020). Die Anzahl an Personen bei Fokusgruppen ist eher geringer, es handelt sich z. B. um fünf bis zehn Personen (Hassan et al. 2020).

Unter den durch qualitative Befragungen in der Wissenschaftskommunikation untersuchten Akteur:innen lassen sich verschiedene Gruppen identifizieren: Expert:innen bzw. Wissenschaftler:innen, Bürger:innen und Vertreter:innen der Zivilgesellschaft sowie praktische Wissenschaftskommunikator:innen und Wissenschaftsjournalist:innen. Auf spezielle Ergebnisse zum Einsatz qualitativer Befragungen bei diesen Gruppen wird im Folgenden eingegangen.

2.1 Qualitative Befragungen von Expert:innen/Wissenschaftler:innen

Eines der Hauptziele, für das qualitative Befragungen wie Leitfadeninterviews oder Fokusgruppen-Diskussionen in der Evaluation von und Forschung zu Wissenschaftskommunikation eingesetzt werden, ist, zu verstehen, welche Kommunikationskanäle Wissenschaftler:innen nutzen. Es interessiert dabei häufig auch, wie Wissenschaftler:innen selbst ihre Wissenschaftskommunikation, z. B. auch über Social Media, einschätzen und reflektieren und welche Motive dahinter liegen (Olesk 2021; Scheu und Schedifka 2018; Sharon und Baram-Tsabari 2020). Hierzu können Dokumente, wie zum Beispiel Screenshots von Social-Media-Threads, eingesetzt werden, um die Erinnerung zu stimulieren und bestimmte Situationen zu rekonstruieren (Sharon und Baram-Tsabari 2020). Auch die Wahrnehmung der eigenen Forschungsthemen in der wissenschaftlichen Forschung im Vergleich zu deren Darstellung in den Medien wird durch qualitative Interviews untersucht (Samuel et al. 2021). Darüber hinaus werden in Studien Leitfadeninterviews von Wissenschaftler:innen kombiniert mit Leitfaden-

interviews von Bürger:innen oder verschiedenen Akteur:innen der Wissenschaft und Wissenschaftspolitik, um unterschiedliche Wahrnehmungen von und Einstellungen zu wissenschaftlichen Themen (Manyweathers et al. 2020) und auch Einschätzungen der Wissenschaftskommunikation selbst (Scheu 2019) herauszuarbeiten.

Auch zur Frage, wie Wissenschaftler:innen bestimmte wissenschaftliche Themen kommunikativ konstruieren, werden qualitative Methoden eingesetzt. Lüthje und Thiele (2018) untersuchen mittels offener Gruppendiskussionen, wie innerhalb der Wissenschaft das Thema und der Begriff Nachhaltigkeit konstruiert und eingeschätzt wird. Mittels Fokus-Gruppendiskussionen wird auch der Frage nachgegangen, wie Wissenschaftler:innen unterschiedlicher fachlicher Disziplinen ihr öffentliches, kommunikatives Engagement einschätzen (Ho et al. 2020).

Als Expert:innen für ein Thema sind aber nicht immer nur Wissenschaftler:innen selbst gefragt, sondern auch Expert:innen in der praktischen Anwendung von wissenschaftlichen Erkenntnissen. So hat Asplund (2020) mittels Fokusgruppen untersucht, wie Landwirt:innen und Vertreter:innen von landwirtschaftlichen NGOs ein Online-Spiel zum Thema Landwirtschaft und Klimawandel evaluieren.

2.2 Qualitative Befragungen von Bürger:innen und der Zivilgesellschaft

Ein weiteres Feld der Wissenschaftskommunikationsforschung, in dem qualitative Befragungen eingesetzt werden, befasst sich mit der Frage, wie Bürger:innen Wissenschaft oder bestimmte wissenschaftliche Themen wahrnehmen, was die dahinterliegenden Gründe sind und zu welchen Handlungen dies führt. So untersuchen zum Beispiel Carmichael et al. (2020) mittels Tiefeninterviews, inwieweit tradierte Narrative die Wahrnehmung von extremen Überflutungen beeinflussen können. Auch in der Forschung zur Klimawandelkommunikation werden qualitative Leitfadeninterviews häufig eingesetzt, um die Wahrnehmungen und Einstellungen zum Klimawandel in verschiedenen Bevölkerungsgruppen und unterschiedlichen Ländern zu verstehen (Mahl et al. 2020; Sarathchandra und Haltinner 2020). Nicht mit einzelnen Bürger:innen, sondern mit Vertreter:innen von zivilgesellschaftlichen Organisationen beschäftigen sich Llorente et al. (2021), die mithilfe von halbstandardisierten Interviews untersuchen, wie diese Organisationen ihren Einbezug in wissenschaftliche Forschung einschätzen.

Zur Analyse der Wahrnehmung von und Einstellung zu wissenschaftlichen Themen werden auch Fokus-Gruppen-Interviews als Form der qualitativen Befragung eingesetzt (Hassan et al. 2020). Gruppendiskussionen kommen ebenso zum Einsatz, wenn es um die Evaluation von bestimmten Formen der Wissenschaftskommunikation selbst durch Bürger:innen geht, wie zum Beispiel von Cititzen Science-Projekten (Rögener und Wormer 2020).

Außerdem lassen sich bei der Analyse der Wahrnehmung von Wissenschaft durch Bürger:innen qualitative Befragungen nicht nur als ergänzende Vertiefung von repräsentativen, standardisierten Befragungen einsetzen (Critchley et al. 2020), sondern auch in Kombinationen mit anderen qualitativen Erhebungsmethoden. So kombinieren Koch et al. (2020) halbstandardisierte Leitfadeninterviews mit der qualitativen Tagebuchmethoden, um herauszufinden, wo Bürger:innen im Alltag Wissenschaft begegnet und wie sie dies einordnen und bewerten.

2.3 Qualitative Befragungen von praktischen Wissenschaftskommunikator:innen und Journalist:innen

Qualitative Befragungen werden in der Forschung auch eingesetzt, um zu eruieren, wie Wissenschaftskommunikator:innen Wissenschaftskommunikation evaluieren (Navarro und McKinnon 2020), warum sich Wissenschaftskommunikator:innen für ein Projekt oder Format der Wissenschaftskommunikation engagieren und wie sie das Ergebnis des Projekts oder des Events bewerten (Almeida und Vaz Bevilaqua 2021). Auch hier werden häufig halbstrukturierte Interviews eingesetzt. So wird unter anderem untersucht, wie die Zusammenarbeit zwischen Wissenschaftler:innen und Künstler:innen bei einem Theaterstück in einem Science Museum funktioniert (Almeida und Vaz Bevilaqua 2021), wie der Wirkungskreis von Wissenschaftskommunikationstrainer:innen aussieht (Dudo et al. 2021) oder warum Universitäten Pressemitteilungen an lokale Presseklubs senden (Koso 2021).

Auch zur Erforschung der aktuellen Praktiken im Wissenschaftsjournalismus und von Einschätzungen von Wissenschaftsjournalist:innen werden qualitative Befragungen genutzt (Burch 2021). So lässt sich beispielsweise analysieren, wie Wissenschaftsjournalist:innen die Wissenschaftsberichterstattung evaluieren und speziell die Präsentation von Unsicherheit von wissenschaftlichen Ergebnissen in den Medien wahrnehmen (Guenther et al. 2015) oder wie sie die Zusammenarbeit mit Wissenschaftler:innen bei bestimmten Formaten der Wissenschaftskommunikation bewerten (MacGregor und Cooper Amanda 2020).

Im Feld der strategischen Kommunikation werden qualitative Befragungen eingesetzt, um die strategische Kommunikation von Organisationen im Bereich Wissenschaft bzw. deren Vertreter:innen zu analysieren (Scheu 2019; VanDyke und King 2020).

Insgesamt zeigt sich, dass qualitative Befragungen in der Wissenschaftskommunikationsforschung durchaus verbreitet sind. Es werden vor allem halbstandardisierte Leitfadeninterviews sowie Fokusgruppen-Interviews mit Wissenschaftler:innen selbst, Bürger:innen, Wissenschafskommunikator:innen und Wissenschaftsjournalist:innen durchgeführt. Qualitative Befragungen dieser Personengruppen können ein tiefergehendes Verständnis hervorbringen, wie diese Personen Wissenschaft sowie spezielle Formen und Formate der Wissenschaftskommunikation wahrnehmen und bewerten. Damit sind qualitative Befragungen eine fruchtbare Methode zur Evaluation von Wissenschaftskommunikation.

3 Fazit: Potenziale qualitativer Befragungen im Kontext der Evaluation von Wissenschaftskommunikation

In diesem Beitrag haben wir eingangs auf methodologischer Ebene verschiedene Formen der qualitativen Befragung vorgestellt, die im Kontext von Evaluationen eingesetzt werden können. Anhand von Beispielen aus der Wissenschaftskommunikationsforschung wurden die Einsatzfelder qualitativer Befragungen in diesem Forschungsfeld aufgezeigt, und wir haben veranschaulicht, wie und mit welchen Zielen das methodische Instrument praktisch eingesetzt wird. Aus den methodologischen Überlegungen und dem Überblick über das Forschungsfeld lassen sich Potenziale für die Evaluationsforschung im Bereich Wissenschaftskommunikation ableiten.

Qualitative Befragungen bieten sich immer dann an, wenn es wichtig erscheint, möglichst offen die Perspektiven, Bewertungen und Einschätzungen bestimmter Zielgruppen zu erheben. Insbesondere bei der Entwicklung neuer Formate von Wissenschaftskommunikation bieten sich daher narrative Befragungen an, zum Beispiel als Einzelinterviews oder in Fokusgruppen. Zur Weiterentwicklung und begleitenden Evaluation bestimmter Wissenschaftskommunikationsformate über die Zeit schlagen wir stärker strukturierte Varianten der qualitativen Befragung vor. Insbesondere Leitfadeninterviews mit Rezipierenden oder Expert:inneninterviews mit spezifischen Stakeholder:innen (z. B. Akteur:innen aus der Hochschulkommunikation, Journalist:innen, Forscher:innen) erscheinen uns hierbei vielversprechend. Denkbar sind außerdem

DELPHI-Studien, mit deren Hilfe Trends und zukünftige Herausforderungen strukturiert und erfasst werden können. Die stärker strukturierten Interviewformate ermöglichen die Vergleichbarkeit im Zeitverlauf und damit auch die Beobachtung von erwünschten und unerwünschten Entwicklungen sowie von Auswirkungen implementierter Veränderungen. Hierbei bieten qualitative Verfahren über die bewusste Auswahl von Befragten (Theoretical Sampling) die Möglichkeit, Sichtweisen unterschiedlicher Akteur:innengruppen zu erfassen und zu vergleichen. So können widersprüchliche Erwartungen, unterschiedliche Interessen und Ansprüche differenziert betrachtet und Wissenschaftskommunikationsformate entsprechend weiterentwickelt werden.

Darüber hinaus sehen wir für die Evaluation bereits etablierter Formate der Wissenschaftskommunikation qualitative Befragungen als sinnvolle Ergänzung oder auch Fundierung von quantitativen Erhebungen. Qualitative Befragungen können dazu dienen, quantitative Evaluationsmethoden zu konstruieren und diese so empirisch fundieren. Sie können aber auch dazu dienen, geschlossene quantitative Formate qualitativ zu öffnen und statische Evaluationsmethoden auf diese Weise dynamischer zu gestalten. Schließlich besteht ein weiterer Mehrwert darin, im Anschluss an quantitative Evaluationsverfahren unerwartete Antworten, Auffälligkeiten oder Widersprüche im Datenmaterial gezielt und vertiefend zu bearbeiten.

Literatur

Almeida C, Vaz Bevilaqua D (2021) The collaboration in the production of Life of Galileo in a science museum in Rio de Janeiro. J Sci Commun 20:1–16. https://doi.org/10.22323/2.20020201

Asplund T (2020) Credibility aspects of research-based gaming in science communication–the case of The Maladaptation Game. J Sci Commun 19:1–20. https://doi.org/10.22323/2.19010201

Blöbaum B, Nölleke D, Scheu AM (2016) Das Experteninterview in der Kommunikationswissenschaft. In: Averbeck-Lietz S, Meyen M (Hrsg) Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft. Springer Fachmedien Wiesbaden, Wiesbaden, S 175–190

Burch E (2021) A sea change for climate refugees in the south pacific: how social media – not journalism – tells their real story. Environ Commun 15:250–263. https://doi.org/10.1080/17524032.2020.1821742

Carmichael C, Danks C, Vatovec C (2020) Assigning blame: how local narratives shape community responses to extreme flooding events in detroit, michigan and waterbury, vermont. Environ Commun 14:300–315. https://doi.org/10.1080/17524032.2019.1659840

Critchley C, Wiersma M, Lipworth W, Light E, Dive L, Kerridge I (2020) Examining diversity in public willingness to participate in offshore human biobanking: an Australian mixed methods study. Public Underst Sci 29:757–769. https://doi.org/10.1177/0963662520948034

Dudo A, Besley JC, Yuan S (2021) Science communication training in North America: preparing whom to do what with what effect? Sci Commun 43:33–63. https://doi.org/10.1177/1075547020960138

Flick U (2007) Qualitative Sozialforschung. Eine Einführung. Rowohlt, Reinbek bei Hamburg

Guenther L, Froehlich K, Ruhrmann G (2015) (Un)Certainty in the news. J & Mass Commun Qtly 92:199–220. https://doi.org/10.1177/1077699014559500

Häder M, Häder S (2014) DELPHI-Befragung. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien Wiesbaden, Wiesbaden, S 587–592

Hassan L, Dalton A, Hammond C, Tully MP (2020) A deliberative study of public attitudes towards sharing genomic data within NHS genomic medicine services in England. Public Underst Sci 29:702–717. https://doi.org/10.1177/0963662520942132

Helfferich C (2014) Leitfaden- und Experteninterviews. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien Wiesbaden, Wiesbaden, S 559–574

Ho SS, Looi J, Leung YW, Goh TL (2020) Public engagement by researchers of different disciplines in Singapore: a qualitative comparison of macro- and meso-level concerns. Public Underst Sci 29:211–229. https://doi.org/10.1177/0963662519888761

Jones SH, Elsdon-Baker F, Catto R, Kaden T (2020) What science means to me: understanding personal identification with (evolutionary) science using the sociology of (non)religion. Public Underst Sci 29:579–596. https://doi.org/10.1177/0963662520923110

Koch C, Saner M, Schäfer MS, Herrmann-Giovanelli I, Metag J (2020) "Space means science, unless it's about star wars": a qualitative assessment of science communication audience segments. Public Underst Sci 29:157–175. https://doi.org/10.1177/0963662519881938

Koso A (2021) The press club as indicator of science medialization: how Japanese research organizations adapt to domestic media conventions. Public Underst Sci 30:139–152. https://doi.org/10.1177/0963662520972269

Küsters I (2014) Narratives Interview. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien Wiesbaden, Wiesbaden, S 575–580

Llorente C, Revuelta G, Carrió M (2021) Social participation in science: perspectives of Spanish civil society organizations. Public Underst Sci 30:36–54. https://doi.org/10.1177/0963662520960663

Löblich M (2016) Theoriegeleitete Forschung in der Kommunikationswissenschaft. In: Averbeck-Lietz S, Meyen M (Hrsg) Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft. Springer Fachmedien Wiesbaden, Wiesbaden, S 67–79

Loosen W (2016) Das Leitfadeninterview – eine unterschätzte Methode. In: Averbeck-Lietz S, Meyen M (Hrsg) Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft. Springer Fachmedien Wiesbaden, Wiesbaden, S 1–15

Lüthje C (2016) Die Gruppendiskussion in der Kommunikationswissenschaft. In: Averbeck-Lietz S, Meyen M (Hrsg) Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft. Springer Fachmedien Wiesbaden, Wiesbaden, S 157–173
Lüthje C, Thiele F (2018) „Nachhaltigkeit ist ein Omnibus, in dem jeder mitfahren darf!" – Die Kommunikative Konstruktion von Nachhaltigkeit in der Wissenschaft. In: Hagen L, Lüthje C, Ohser F, Seifert C (Hrsg) Wissenschaftskommunikation: die Rolle der Disziplinen. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden, S 151–176
MacGregor S, Cooper A (2020) Blending research, journalism, and community expertise: a case study of coproduction in research communication. Sci Commun 42:340–368. https://doi.org/10.1177/1075547020927032
Mahl D, Guenther L, Schäfer MS, Meyer C, Siegen D (2020) "We are a bit blind about it": A qualitative analysis of climate change-related perceptions and communication across South African communities. Environ Commun 14:802–815. https://doi.org/10.1080/17524032.2020.1736116
Manyweathers J, Taylor M, Longnecker N (2020) Expertise and communicating about infectious disease: a case study of uncertainty and rejection of local knowledgein discourse of experts and decision makers. J SciCommun 19. https://doi.org/10.22323/2.19040201
Navarro K, McKinnon M (2020) Challenges of communicating science: perspectives from the Philippines. J SciCommun 19:1–21. https://doi.org/10.22323/2.19010203
Olesk A (2021) The types of visible scientists. J SciCommun 20:1–18. https://doi.org/10.22323/2.20020206
Rögener W, Wormer H (2020) Gute Umweltkommunikation aus Bürgersicht. Ein Citizen-Science-Ansatz in der Rezipierendenforschung zur Entwicklung von Qualitätskriterien. M&K Medien & Kommunikationswissenschaft 68:447–474. https://doi.org/10.5771/1615-634X-2020-4-447
Samuel G, Diedericks H, Derrick G (2021) Population health AI researchers' perceptions of the public portrayal of AI: a pilot study. Public Underst Sci 30:196–211. https://doi.org/10.1177/0963662520965490
Sarathchandra D, Haltinner K (2020) Trust/distrust judgments and perceptions of climate science: a research note on skeptics' rationalizations. Public Underst Sci 29:53–60. https://doi.org/10.1177/0963662519886089
Scheu AM (2016) Grounded Theory in der Kommunikationswissenschaft. In: Averbeck-Lietz S, Meyen M (Hrsg) Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft. Springer Fachmedien, Wiesbaden, S 81–94
Scheu AM (2019) Between offensive and defensive mediatization. An exploration of mediatization strategies of German science-policy stakeholders. J SciCommun 18:1–20. https://doi.org/10.22323/2.18030208
Scheu AM, Schedifka T (2018) Wissenschaftskommunikation im Netz. Eine explorative Studie zur Nutzung webbasierter sozialer Kommunikationskanäle. In: Hagen L, Lüthje C, Ohser F, Seifert C (Hrsg) Wissenschaftskommunikation: die Rolle der Disziplinen. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden, S 177–212
Scholl A (2009) Die Befragung. UTB, Stuttgart
Sharon AJ, Baram-Tsabari A (2020) The experts' perspective of "ask-an-expert": an interview-based study of online nutrition and vaccination outreach. Public Underst Sci 29:252–269. https://doi.org/10.1177/0963662519899884

Strübing J (2008) Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung des Verfahrens der empirisch begründeten Theoriebildung. Springer VS, Wiesbaden

VanDyke MS, King AJ (2020) Dialogic communication practices of water District officials: insights from practitioner interviews. Environ Commun 14:147–154. https://doi.org/10.1080/17524032.2019.1705365

Vogl S (2014) Gruppendiskussionen. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien Wiesbaden, Wiesbaden, S 581–586

Julia Metag ist Professorin für Kommunikationswissenschaft am Institut für Kommunikationswissenschaft der Westfälischen Wilhelms-Universität Münster. In ihrer Forschung beschäftigt sie sich mit Wissenschaftskommunikation und politischer Kommunikation. Sie hat umfangreich zu Publikumssegmenten in der Wissenschaftskommunikation, Nutzung und Wirkung von Wissenschaftskommunikation sowie visueller Wissenschaftskommunikation publiziert.

Andreas M. Scheu ist wissenschaftlicher Leiter der Transfer Unit Wissenschaftskommunikation an der Berlin-Brandenburgischen Akademie der Wissenschaften, Privatdozent am Institut für Kommunikationswissenschaft der Westfälischen Wilhelms-Universität Münster und Sprecher der Fachgruppe Wissenschaftskommunikation der Deutschen Gesellschaft für Publizistik und Kommunikationswissenschaft. Andreas forscht schwerpunktmäßig in den Bereichen Wissenschaftskommunikation und Medialisierung und veröffentlicht Beiträge zur Weiterentwicklung qualitativer Methoden.



Praxisbeitrag: Qualitative Befragungen zur Evaluation von Wissenschaftskommunikation am Beispiel des Wissenschaftsvarietés Glitzern & Denken

Imke Hedder, Ricarda Ziegler, Bonnie Dietermann und David Ziegler

Zusammenfassung

Wissenschaft und Kunst zu verbinden, zeichnet inzwischen viele Formate der Wissenschaftskommunikation aus. Wie Künstler:innen und Wissenschaftler:innen durch ihr Zusammenwirken wahrgenommen werden, wurde in der Evaluation des Wissenschaftsvarietés *Glitzern & Denken* beleuchtet. Leitfadengestützte Vorher-Nachher-Interviews mit Zuschauenden dienten dem Zweck, Zielgruppen des Formats näher kennenzulernen, ihre Vorstellungen von Wissenschaft und Kunst sowie ihre Programmbewertung zu erheben. Auf methodischer Ebene zeigt das Fallbeispiel auf, wie die Umstellung des ursprünglich reinen Präsenzformats zu einem digitalen

I. Hedder (✉) · R. Ziegler
Wissenschaft im Dialog, Berlin, Deutschland
E-Mail: imke.hedder@w-i-d.de

R. Ziegler
E-Mail: ricarda.ziegler@w-i-d.de

B. Dietermann · D. Ziegler
Museum für Naturkunde – Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Deutschland
E-Mail: bonnie.dietermann@mfn.berlin

D. Ziegler
E-Mail: david.ziegler@mfn.berlin

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*,
https://doi.org/10.1007/978-3-658-39582-7_8

Angebot die Evaluationsplanung veränderte und verdeutlicht das Erkenntnispotenzial, das aus wenigen, sorgfältig vorbereiteten Befragungen gewonnen werden kann. Auf inhaltlicher Ebene geben die Ergebnisse Anlass zur Reflexion, wie die Wahrnehmung von Wissenschaftler:innen durch ihre Präsentation und die Moderation in Formaten der Wissenschaftskommunikation geprägt wird.

Im Rahmen des Projekts *Glitzern & Denken – das Wissenschaftsvarieté* wurden von 2019 bis 2022 am Museum für Naturkunde Berlin (MfN) zu verschiedenen naturwissenschaftlichen Themen Varieté-Abende konzipiert und umgesetzt. Wissenschaft im Dialog (WiD) begleitete das Projekt u. a. als Evaluationspartner.[1] Im folgenden Beitrag wird skizziert, wie die im Herbst 2020 dreimalig aufgeführte und digital übertragene Show mit dem Titel *Schleimig!* zum Thema Weichtiere (*Mollusca* – also Schnecken, Muscheln und Tintenfische) durch qualitative Vorher-Nachher-Interviews evaluiert wurde.

1 Das Projekt Glitzern & Denken und seine Ziele

Die Idee hinter *Glitzern & Denken* war es, Performance-Kunst mit Wissenschaft zu verbinden, inspiriert von den künstlerischen und wissenschaftlichen Salons des 18. und 19. Jahrhunderts. Hierfür entwickelten Künstler:innen in Zusammenarbeit mit Wissenschaftler:innen ein Varieté zu wissenschaftlichen Themen. Fakten aus der aktuellen Forschung sowie interessante Geschichten aus Naturwissenschaften und Kulturgeschichte wurden gemeinsam mit Akrobatik und thematisch abgestimmter Musik präsentiert.[2]

Das Format zielte auf eine diversere öffentliche Kommunikation über Wissenschaft und insbesondere die naturwissenschaftliche Forschung am MfN ab. Das Format versprach sowohl eine intellektuelle als auch emotionale Ansprache des Publikums. Dadurch wurde angestrebt, einerseits bestehende Zielgruppen zu erreichen, die bereits wissenschaftskommunikative Angebote (speziell auch des MfN) wahrnehmen. Ihnen sollten neue Facetten des Hauses und dessen

[1] Umsetzungspartner von *Glitzern & Denken* und verantwortlich für die inhaltliche Ausgestaltung und künstlerische Durchführung war das *Ensemble Salon Fähig* unter Leitung von Ines Theileis. Gefördert wurde das Projekt durch die LOTTO-Stiftung Berlin.

[2] Weitere Informationen sind auf der Projektwebsite zu finden: https://www.museumfuernaturkunde.berlin/de/museum/ausstellungen/glitzern-denken-das-wissenschaftsvariete.

Forschung eröffnet sowie ein Kennenlernen der Mitarbeitenden und Forschenden der wissenschaftlichen Sammlungen ermöglicht werden. Andererseits sollten durch das Varieté-Format auch Personen angesprochen werden, die bisher nicht zu den erreichten Zielgruppen des MfN gehören – insbesondere Menschen, die zwar ein bestehendes Interesse für Formate der Performance-Kunst haben, aber eine gewisse Distanz zu naturwissenschaftlichen Themen aufweisen. Mit dem Wissenschaftsvarieté wollte man in der neuen Zielgruppe ein erstmaliges Interesse an Wissenschaftskommunikationsangeboten erwecken. Darüber hinaus galt für beide Zielgruppen das Ziel, mit dem Format ein positives Erlebnis zu schaffen und eine neuartige Auseinandersetzung mit Wissenschaft und Forschung zu ermöglichen, die ihre Perspektiven auf ebendieses Feld erweitern.

2 Evaluationsinteressen und Verortung der Evaluation im Projekt Glitzern & Denken

Von den im vorherigen Abschnitt präsentierten qualitativen Projektzielen abgeleitet, sollte in der Evaluation der *Schleimig!*-Shows im Herbst 2020 insbesondere den folgenden Fragen nachgegangen werden:

- Erreicht die Show neben den bereits bestehenden Zielgruppen des MfN auch neue Zielgruppen und kann sie diese für das Format gewinnen?
- Welche Bilder von Kunst und Wissenschaft sowie Wissenschaftler:innen und Künstler:innen bringt das Publikum (bzw. die unterschiedlichen Zielgruppen) mit? Werden diese durch die Shows beeinflusst?
- Wie werden die Varieté-Shows vom Publikum (bzw. von den unterschiedlichen Zielgruppen) wahrgenommen? Wie wird das Programm der Shows bewertet und wie ließe es sich aus Publikumssicht noch verbessern?

Die Beantwortung dieser Evaluationsfragen im laufenden Projekt sollte Anpassungen im Prozess ermöglichen und damit die Verwirklichung der übergeordneten Projektziele über die gesamte Projektlaufzeit unterstützen.

Die Varieté-Veranstaltungen von *Glitzern & Denken* waren ursprünglich als Live-Shows vor einem Publikum von ca. 150 Personen pro Abend im *Experimentierfeld für Partizipation und Offene Wissenschaft* des MfN geplant. Aufgrund der anhaltenden COVID-19-Pandemie wurde das Format digital umgesetzt und im Herbst 2020 live auf den YouTube- und Facebook-Kanälen des Museums vor einem Online-Publikum gestreamt. Die ursprünglich

geplanten Evaluationsmethoden (u. a. halbstandardisierte Vorher-Nachher-Interviews vor Ort und teilnehmende Beobachtungen) konnten nicht wie geplant durchgeführt werden und wurden auf diese Gegebenheiten angepasst: Damit fokussierte sich die Evaluation neben der Analyse der YouTube-Daten auf eine Vor- und Nachbefragung der Zuschauenden über weitestgehend standardisierte Online-Fragebögen und eine inhaltsanalytische Auswertung der YouTube- und Facebook-Kommentare. Aufgrund geringer Fallzahlen konnten allerdings keine belastbaren Ergebnisse generiert werden, weshalb zusätzlich eine nachträgliche Datenerhebung mittels qualitativer, leitfadengestützter Vorher-Nachher-Interviews angeschlossen wurde, bei der die speziell hierfür rekrutierten Interviewpartner:innen eine Aufzeichnung der *Schleimig!*-Shows über YouTube (als Stimulus) sahen. Diese Datenerhebungen stehen im Fokus dieses Beitrags.

Dennoch lieferten die Daten aus Online-Fragebögen, YouTube-Analytics und den Kommentarspalten wertvolle Informationen für die zusätzliche Datenerhebung über leitfadengestützte Vorher-Nachher-Interviews und ihre qualitative Auswertung; beispielsweise Anhaltspunkte für Kategorien der Assoziationen mit Wissenschaft und Kunst, die aus den Antworten der Befragten auf offene Fragen im Online-Fragebogen gewonnen wurden. Zusätzlich lieferten die Ergebnisse der standardisierten Erhebung des Freizeitverhaltens und der Soziodemografie der Online-Befragten erste Hinweise, dass sich das Show-Publikum nicht zwangsläufig trennscharf in die MfN-bekannte Zielgruppe und eine neu erschlossene, eher an Performance-Kunst statt an Wissenschaften interessierte Zielgruppe einordnen ließ.

Offene Fragen für die Evaluation bezüglich der Bewertung der einzelnen Programmpunkte hinterließen die Schwankungen in den Zuschauerzahlen des Livestreams. Ob sich hohe Absprungzahlen durch einen aktuell laufenden Programmpunkt erklären ließen oder pragmatisch begründet waren, weil die Zuschauer:innen zu Hause etwa zu bestimmten Uhrzeiten anderen Verpflichtungen nachkommen mussten (beispielsweise Kinder ins Bett zu bringen), blieb unklar.

3 Durchführung der qualitativen Befragungen

Aufbauend auf diesen ersten Ergebnissen der Datenerhebungen während der Live-Shows wurden die qualitativen Befragungen in Form von leitfadengestützten Vorher-Nachher-Interviews im Zeitraum vom 18. Juni 2021 bis 13. Juli 2021 durchgeführt und anschließend ausgewertet. Diese Art der Datenerhebung erlaubte es, tiefe Einblicke in das „subjektive Erleben" des Publikums

zu gewinnen (Döring und Bortz 2016, S. 356), Assoziationsmuster mit Wissenschaft und Kunst durch Nachfragen näher zu erörtern und durch offene Fragen auf unerwartete Blickwinkel zu stoßen (siehe auch Metag und Scheu in diesem Band). Die Gesprächssituation eignete sich auch gut dazu, die Zielgruppen eines neuen Formats wie *Glitzern & Denken* besser kennenzulernen und zu verstehen.

3.1 Vorbereitung der Interviews

Die Auswahl der Interviewpartner:innen geschah auf Basis eines Screening-Fragebogens, der mit einem Aufruf zur Teilnahme an der Evaluation über Social-Media-Kanäle geteilt wurde.[3] Über diese kurze Online-Befragung konnten Personen ihr Interesse an einer Teilnahme signalisieren und ihre Kontaktdaten hinterlassen. Ziel dieses Vorgehens war es, Interviewpartner:innen zu identifizieren, welche die beiden Zielgruppen der Show im ähnlichen Verhältnis abbildeten. Die Unterscheidung dieser Zielgruppen wurde an der Frage festgemacht, wie häufig sie im Regelfall[4] im Laufe eines Jahres bestimmte Freizeitaktivitäten wahrnehmen, die entweder ein Interesse an künstlerischen Darbietungen oder an wissenschaftlichen Themen signalisieren. Darüber hinaus wurden mit dem Screening-Fragebogen soziodemografische Angaben sowie die bisherige Kenntnis des Formats des Wissenschaftsvarietés und des Projekts *Glitzern & Denken* erhoben.

Von den insgesamt 42 Interessierten wurden zwei Kandidat:innen ausgewählt, die ein stärkeres Interesse an Kunst signalisierten und gleichzeitig ein geringeres Interesse an wissenschaftlichen Themen; zwei weitere, deren Freizeitverhalten auf ein größeres Interesse an Wissenschaft und ein geringeres Interesse an Kunst deutete; und schließlich zwei Personen, die keine klare Zuordnung auf Basis des Freizeitverhaltens zuließen, aber aufgrund ihres Alters und formalen Bildungsniveaus interessante, ergänzende Perspektiven versprachen. Die Auswahl fiel bewusst auf Personen, die *Glitzern & Denken* noch nicht kannten. Dies erlaubte einen unbeeinflussten Einblick in Erwartungen und Assoziationen und bot damit

[3] Der Aufruf wurde verbreitet vom MfN, WiD und Multiplikator:innen mit Berührungspunkten zur Kunstszene über Facebook, Instagram und Twitter.

[4] Es wurde nach dem Freizeitverhalten in der Zeit vor der COVID-19-Pandemie gefragt. Dass die Erinnerung nach knapp einem Jahr etwas verzerrt sein kann, wurde berücksichtigt, allerdings zum Zweck der Identifikation genereller Tendenzen als vertretbar bewertet.

ideale Voraussetzungen für die Vorbefragung, um schlussendlich Vergleiche ziehen zu können.

Inhaltlich wurden die qualitativen Befragungen vorbereitet, indem basierend auf den Erkenntnisinteressen der Evaluation Themenblöcke entwickelt, anschließend konkrete Interviewfragen innerhalb der Themenblöcke formuliert und schließlich jene Fragen den Vorher- und/oder Nachher-Interviews zugeordnet wurden. Anschließend wurden die Themenblöcke in einer für die Interviewpartner:innen möglichst intuitiven Reihenfolge[5] angeordnet. In den Themenblöcken wurden die (in der Regel ca. vier) Fragen so geordnet, dass zunächst ein allgemeiner Eindruck eingeholt und anschließend näher ins Detail gegangen wurde. Bezüglich der konkreten Formulierung der Fragen wurden möglichst ausgewogene Fragestellungen gewählt, die weder eine positive oder negative Antwort nahelegten, und darüber hinaus jegliche suggestive Formulierungen vermieden.

Die zehn- bis fünfzehnminütigen Vorher-Interviews, welche mit den konkreten Anweisungen für die Rezeption verbunden wurden, beinhalteten die nähere Abfrage des Freizeitverhaltens (um die Zielgruppenzuordnung zu überprüfen), Erwartungen an ein Wissenschaftsvarieté und die aktuellen Vorstellungen von Wissenschaft und Kunst sowie Wissenschaftler:innen und Künstler:innen. Anschließend hatten die Interviewpartner:innen drei bis fünf Tage Zeit, um sich die Aufzeichnung der vergangenen Show *Schleimig!* auf YouTube anzusehen. Es wurde darum gebeten, die Rezeption so natürlich wie möglich zu gestalten – wer die Show in Gesellschaft anschauen wollte, durfte dies tun, auch kurze Pausen waren erlaubt. Die zweiten Interviews dauerten zwischen 20 und 30 Minuten. Zu Beginn wurden hier im gleichen Wortlaut die Assoziationen zu Wissenschaft und Kunst sowie zu Wissenschaftler:innen und Künstler:innen abgefragt. Erst danach widmete sich das Gespräch der Besprechung des Programms, beginnend mit allgemeinen Eindrücken und dem Abgleich mit zuvor formulierten Erwartungen und weiterführend mit der Besprechung konkreter Programmpunkte. Abschließend wurden Verbesserungsvorschläge gesammelt.

3.2 Vorgehen bei der Datenauswertung

Die Interviews wurden aufgezeichnet und überwiegend stichwortartig transkribiert. Besonders prägnante Zitate und Aussagen bei der Abfrage von Assoziationen

[5] So wurden beispielsweise leicht beantwortbare Fragen zu den Interviewpartner:innen selbst an den Anfang gesetzt, um den Interviewpartner:innen den Einstieg in die Befragung möglichst einfach zu gestalten.

wurden jedoch wortgetreu übernommen und höchstens grammatikalisch geglättet. Die Transkripte wurden in einem vorbereiteten Analyseraster angelegt, um direkte Vergleiche zwischen den Zielgruppen sowie Vorher-Nachher-Vergleiche der Einzelpersonen vornehmen zu können.

Die Datenauswertung folgte einem qualitativen Forschungsansatz, der sich besonders dafür eignet, neue Muster auf Basis des vorliegenden Materials zu entdecken und Theorien zu entwickeln (Döring und Bortz 2016, S. 26). So wurde beispielsweise eine induktive Kategorienbildung (siehe auch Metag und Scheu in diesem Band) des Freizeitverhaltens vorgenommen, um nach unterschiedlichen Mustern zwischen den Interviewpartner:innen zu suchen. Hierfür wurden zunächst alle Antworten betrachtet, um erste mögliche Kategorien der Freizeitgestaltung zu sammeln. Anschließend wurden einige Kategorien thematisch zusammengelegt, um dann nochmals alle Antworten dieser finalen Kategorisierung folgend zu kodieren.

Die Auswertung zur Beantwortung der weiteren Evaluationsfragen folgte hingegen einer deduktiven Logik (siehe auch Metag und Scheu in diesem Band). Für die Analyse der Programmbewertung lag ein vorstrukturiertes Raster vor, in dem Aussagen in neutrale Kommentare, positives und negatives Feedback sowie Verbesserungsvorschläge unterteilt wurden. Die Verbesserungsvorschläge wurden vorab in technische Anmerkungen (z. B. zum Bühnenbild, den Lichtverhältnissen oder der Kameraführung), inhaltliche Anmerkungen (z. B. zu gewünschten Gesprächsthemen oder der Verständlichkeit) und strukturelle Anmerkungen (z. B. zu der Moderation oder der Länge bestimmter Programmpunkte) unterteilt.

Um die möglichen Wirkungen der Show auf Vorstellungen von Wissenschaft und Kunst sowie von Wissenschaftler:innen und Künstler:innen zu untersuchen, wurden die Interviewpartner:innen im Vorfeld und im Nachgang der Show gefragt, welche Assoziationen die Begriffe *Wissenschaft* und *Kunst* in ihnen auslösen und welche Eigenschaften und Kompetenzen sie mit *Wissenschaftler:innen* sowie *Künstler:innen* verbinden. Die Antworten wurden in ein Kodierschema eingeordnet, das dem Wissenschaftsbarometer (Ziegler et al. 2018, S. 10 f.) entlehnt ist, aber für die vorgestellten Erkenntnisinteressen angepasst und ergänzt wurde (siehe Tab. A.1 und A.2). Im Vorher-Nachher-Vergleich ließ sich erkennen, welche Antworten wiederholt, ergänzt oder nicht wieder aufgegriffen wurden – und dementsprechend, welche Vorstellungen nach der Show in den Vordergrund traten und welche in den Hintergrund rückten.

4 Ergebnisse der qualitativen Befragungen

Zunächst wird auf die Evaluationsergebnisse zur Erreichung der Zielgruppen von *Glitzern & Denken* eingegangen, da ihre Unterscheidung auch für die Beantwortung der weiteren Evaluationsfragen relevant ist. Die Auswertung des Screening-Fragebogens, dessen Ergebnisse zur Rekrutierung und Auswahl der Interviewpartner:innen genutzt wurden, deutete bereits an, dass basierend auf den Fragen zum Freizeitverhalten nur in wenigen Fällen eine trennscharfe Zuordnung in eine der beiden eingangs skizzierten Zielgruppen des Projekts möglich war. Die Vorher-Nachher-Interviews verstärkten diesen Eindruck, denn alle Interviewpartner:innen nannten ähnliche Formen der Freizeitgestaltung, von Sport und Bewegung über soziale Aktivitäten bis hin zu Kultur und Unterhaltung. Randbemerkungen der Interviewten zu dieser Frage zeigten, dass das Freizeitverhalten stärker von der aktuellen Lebenssituation abzuhängen schien als von einer Neigung für Kunst und/oder Wissenschaft: von zeitlichen Ressourcen, Verantwortlichkeiten wie Kinderbetreuung oder Studium oder auch davon, ob man in der Stadt oder ländlich wohnt. Auch den beruflichen Hintergrund gilt es hier zu beachten. Beide Interviewpartner:innen, die basierend auf ihren Angaben zum Freizeitverhalten bei der Rekrutierung in die erste, mit Wissenschaftskommunikationsangeboten (des MfN) bekannte Zielgruppe eingeordnet wurden, arbeiteten selbst im akademischen Bereich, weshalb sie häufig wissenschaftsbezogene Veranstaltungen besuchen – dies stellt allerdings nicht zwangsläufig ein Freizeitvergnügen dar. Personen, die bisher wenig Berührungspunkte mit Wissenschaft hatten, suchten stattdessen gezielter nach neuen, noch ungewohnten Einblicken in die Wissenschaft. Die fehlende Trennschärfe der ursprünglichen Zielgruppeneinteilung zeigte sich auch bei den Erwartungen an ein Wissenschaftsvarieté. Der Wunsch, Neues zu lernen, unterhalten zu werden und eine neue Perspektive auf Wissenschaft zu gewinnen, tauchte bei Interviewpartner:innen aller Gruppen auf. Ähnlich verhielt es sich mit den Vorstellungen von Wissenschaft und Kunst, die kaum Anhaltspunkte für strukturelle Unterschiede zwischen den Zielgruppen lieferten.

Doch auch ohne die Unterteilung in verschiedene Gruppen lieferte die Auswertung der Vorstellungen von Wissenschaft und Kunst interessante Erkenntnisse. Die Kodierungen der entsprechenden Assoziationen der Interviewpartner:innen fielen im Vorher-Nachher-Vergleich sehr unterschiedlich aus und boten – mit einer Ausnahme im Bereich Kunst – neben Begriffen für die eigens formulierten Kategorien auch Beispiele für alle Kategorien, die aus dem Wissenschaftsbarometer übernommen wurden (siehe Tab. A.1 und A.2). Auch wenn aufgrund der geringen Anzahl von Interviewpartner:innen und der damit überschaubaren Anzahl

an Nennungen keine signifikanten Unterschiede und Veränderungen zwischen den Kategorien präsentiert werden können, sollen im Folgenden dennoch einige erkenntnisreiche deskriptive Einblicke beispielhaft wiedergegeben werden.

Assoziationen mit Wissenschaft und Kunst:

- Vor der Show assoziierten Interviewpartner:innen mit Wissenschaft und Forschung vorrangig wissenschaftliche Disziplinen aus den Naturwissenschaften und persönliche Berührungspunkte sowie (wertende) Wahrnehmungen, wie „kann ich nicht", „Elfenbeinturm", aber auch positive Assoziationen wie „bunt" und „interessant". Im Nachher-Interview nahmen Beispiele und Wertungen ab und Anmerkungen zur Arbeitsweise in der Wissenschaft oder ihrer Erkenntnisorientierung nahmen zu.
- Wissenschaftler:innen wurden vor der Show positiv und idealisiert beschrieben, nur ihre Kommunikationskompetenzen wurden in drei Interviews kritisch kommentiert. Nach der Show wurden Begriffe zu ihrer Integrität vermehrt genannt, die Kritik an ihrer Kommunikation trat nicht mehr auf.
- Wissenschaftler:innen und Künstler:innen galten nicht als Gegensätze. Beiden wurde eine Hartnäckigkeit und Offenheit zugeschrieben. Unterschiede zeigten sich in sozialen Aspekten: Mit Künstler:innen wurde im Gegensatz zu Wissenschaftler:innen eine Nähe und Kommunikationskompetenz verbunden, die nach der Show noch verstärkt genannt wurde. Wissenschaftler:innen wurden nach der Show zwar nicht mehr mit begrenzten Kommunikationskompetenzen assoziiert, aber auch nicht in diesem Maße lobend hervorgehoben. Eine Erklärung hierfür boten Interviewpartner:innen, die als Beispiel die Moderatorin der Show, die Sängerin und Leiterin des Ensembles, nannten. In ihren Augen waren die Künstler:innen die Verbindung zwischen Wissenschaft und Publikum – sie übernahmen die Wissenschaftskommunikation, nicht die Wissenschaftler:innen selbst.

Programmbewertung und Feedback:

- Die Show wurde insgesamt – mit Blick auf die Struktur, die Inhalte, die wissenschaftlichen Gäste wie auch die künstlerischen Darbietungen – positiv wahrgenommen.
- Auf inhaltlicher Ebene wurden Vorschläge für weitere mögliche Gesprächsthemen (z. B. Einblicke in den Forschungsalltag) zum Show-Motto genannt.
- Auf struktureller Ebene kamen Impulse zur (zeitlichen) Schwerpunktsetzung im Programm, bzgl. des Verhältnisses von Publikumsfragen, wissenschaftlichen Beiträgen und künstlerischen Darbietungen.

- Auf technischer Ebene wurden Tipps für die Ausleuchtung, die Kameraführung und das Bühnenbild gegeben, damit die Show für das Online-Publikum besser am Bildschirm zu verfolgen ist.
- Offene Anmerkungen der Interviewten wiesen außerdem auf nicht-intendierte Wirkungen subtiler, unterbewusster Interaktionen zwischen den Akteur:innen hin: Beispielsweise fielen zwei Interviewpartner:innen Rückbezüge auf vorherige Shows und Proben auf, die sie irritierten.

5 Fazit

Im Verlauf des Projekts *Glitzern & Denken* wurde die qualitative Befragung aufgrund der pandemischen Entwicklungen und ihrer unvorhergesehenen Folgen für die Umsetzung und Evaluation der Shows gewählt. Die Methodik geht, wie jede andere auch, mit einigen methodischen Einschränkungen einher: Die Ergebnisse der Interviews können aufgrund der geringen Fallzahl kein umfassendes, verallgemeinerndes Bild liefern. Zudem sollte erwähnt werden, dass der zeitliche Abstand zwischen den Interviews vor und nach Rezeption der Show teilweise relativ kurz war, weshalb sich einige der Interviewten noch sehr gut an ihre Antworten im Vorher-Interview erinnern konnten und verunsichert waren, ob sie der Konsistenz halber die gleichen Vorstellungen von Wissenschaft und Kunst erneut nennen oder bewusst neue Begriffe einbringen sollten[6]. Es kann nicht ausgeschlossen werden, dass die Erinnerung an das Vorher-Interview und der Wunsch, konsistente Antworten zu geben oder Abwechslung zu liefern, in die Ergebnisse einflossen.

Dabei sind allerdings die vielen Vorteile des qualitativen Zugangs nicht zu unterschlagen: Für ein neues Projektkonzept, das bisher relativ unerforscht ist, bieten Interviews tiefere Einblicke, um Wahrnehmungen und mögliche Wirkungsmechanismen zu erkunden. Insgesamt konnten mit relativ überschaubarem Aufwand für die Evaluation – zwölf Interviews mit einer Dauer zwischen zehn und dreißig Minuten – viele Ergebnisse gewonnen werden, welche die weitere Entwicklung des Projekts auf vielfältige Art unterstützten. Neben den zahlreichen Anmerkungen hinsichtlich der Bildschirmoptimierung der Show für das Online-Publikum und inhaltlicher Impulse stachen vor allem zwei besonders wertvolle Erkenntnisse hervor.

[6] Anmerkungen dieser Art wurden im Transkript kenntlich gemacht. Es wurde dazu ermutigt, einfach die ersten Begriffe zu nennen, die den Personen einfielen – ohne dabei zu berücksichtigen, ob dies zu neuen Begriffen oder Dopplungen führen würde.

Zum einen stellten die Gespräche die Unterteilung in eine wissenschaftsaffine Zielgruppe auf der einen Seite und eine kunstaffine und gleichzeitig wissenschaftsferne Zielgruppe auf der anderen Seite in Frage. Es ergeben sich verschiedene Erklärungsansätze, weshalb sich diese Zielgruppentrennung nicht in der Evaluation bestätigte: Möglicherweise wurden Repräsentant:innen dieser Gruppen im Rahmen der Rekrutierung für die vorgestellte Evaluation nicht erreicht oder die Operationalisierung der Zielgruppen über das jeweilige Freizeitverhalten, die im Screening-Fragebogen und während der Interviews gewählt wurde, war nicht zielführend. Eine weitere Erklärung wäre, dass die dahinter liegenden Annahmen in der Projektkonzeption noch nicht ausgereift waren, um die zentralen Unterscheidungsmerkmale zwischen den gewünschten Zielgruppen klar zu benennen. Die qualitativen Ergebnisse können hierfür nun neue Inspiration liefern, um groben Vorstellungen mehr inhaltliche Tiefe zu verleihen.

Die Ergebnisse der Interviews wiesen dabei auf andere Kriterien für die Unterscheidung von Zielgruppen hin, die in der weiteren Kommunikation und Programmgestaltung im Projekt von Interesse waren, so zum Beispiel die Motive für einen Museumsbesuch und die Auseinandersetzung mit den Inhalten von *Glitzern & Denken*. Dieser Impuls ergab sich aus den Gesprächen mit den bereits im akademischen Bereich Tätigen, deren Motive sich von den Personen fernab des akademischen Systems unterschieden. Die Affinität für künstlerische oder (natur-)wissenschaftliche Inhalte mag sich somit nicht direkt in der Freizeitgestaltung widerspiegeln, sondern im Kontext von *Glitzern & Denken* eher in der Art und Weise, wie man den Inhalten der Varieté-Show allgemein begegnen würde, worin diesbezügliche Neugierde oder Desinteresse begründet sein könnten oder auch, inwieweit man sich in der Lage fühlt, naturwissenschaftliche Themen zu verstehen.

Diese Erkenntnis legt für künftige Projekte nahe, mehr Energie in die Analyse der Zielgruppen des Projekts zu investieren, um ihre Beweggründe und Abgrenzungsmerkmale besser zu identifizieren, kommunikativ gezielter zu adressieren, ihren Bedürfnissen inhaltlich besser zu begegnen und schlussendlich auch evaluativ erfassen zu können, ob diese tatsächlich erreicht wurden. Im Fall der *Glitzern & Denken*-Evaluation boten die Ergebnisse neue Denkanstöße, um in der Evaluation darauffolgender Shows neue Abgrenzungsmerkmale der Zielgruppen zu testen.

Zum anderen zeigten die Interviews unintendierte Effekte des Formats auf: Mit der Konzeption des Formats wurde erhofft, dass durch die Verbindung von Kunst und Wissenschaft nicht nur die wissenschaftlichen Inhalte unterhaltsamer und nahbarer gestaltet werden, sondern auch das Image der Forschenden vom gemeinsamen Bühnenauftritt mit Künstler:innen profitiert, indem ihre sozialen

und kommunikativen Eigenschaften hervorgehoben werden – Aspekte, die nicht als die zentralen Charakteristiken von Wissenschaftler:innen gelten (Ziegler et al. 2018, S. 7). Zwar zeigte sich, dass die Wissenschaftler:innen ein allgemein positives Feedback bekamen und diese offenbar in der Lage waren, das anfangs negative Vorurteil ihrer begrenzten Kommunikationsfähigkeiten in den Augen der Interviewpartner:innen zumindest abzuschwächen, gleichzeitig standen sie jedoch im direkten Vergleich mit den schillernden Persönlichkeiten der Künstler:innen. Der Effekt einer starken Kontrastierung zwischen den Wissenschaftler:innen auf der einen Seite und den Künstler:innen als bühnenerfahrenen, professionellen Kommunikator:innen auf der anderen Seite, sollte nicht unterschätzt werden. Auch kommunikativ talentierte Wissenschaftler:innen könnten in der direkten Gegenüberstellung zu Künstler:innen als verhältnismäßig schwächer kommunizierend wahrgenommen werden. Somit kann der gegenteilige Effekt der eigentlichen Zielsetzung dieser künstlerisch-wissenschaftlichen Veranstaltung eintreten: Wissenschaftler:innen können nicht etwa als nahbare und kommunikative Personen, sondern gegenüber den Künstler:innen als „blass“ und unnahbar erscheinen. Dieser Effekt wurde bei *Schleimig!* gegebenenfalls noch dadurch verstärkt, dass die Künstler:innen in ihrer Performance mit dem Stereotyp des schwer verständlichen Wissenschaftlers oder der „nerdigen“ Wissenschaftlerin spielten, was aus Sicht der künstlerischen Darbietung sicherlich nachvollziehbar ist, für die ultimativen Projektziele von *Glitzern & Denken* jedoch schlussendlich als wenig förderlich eingeschätzt wurde. Im Fall von *Glitzern & Denken* gab dieses Ergebnis den Impuls, auch den Wissenschaftler:innen mehr Möglichkeiten in der Show zu geben, um ihr kommunikatives Talent unter Beweis zu stellen und die Moderation für mögliche unintendierte Effekte zu sensibilisieren. Diese Erkenntnis kann sicher auch für andere Formate, in denen Wissenschaftler:innen auf der Bühne mit Kommunikationsprofis – seien es Künstler:innen oder ausgebildete Moderator:innen – agieren, gewinnbringend sein. Darüber hinaus wären weitere Untersuchungen dazu interessant, wie Wissenschaftler:innen in solchen Veranstaltungen wahrgenommen werden, die bewusst mit kontrastierenden Persönlichkeiten und Klischees spielen.

Die qualitativen Befragungen von Zuschauer:innen der Show beleuchteten Stärken und Fremdwahrnehmungen, aber auch Optimierungspotenziale und mögliche unintendierte Effekte von *Glitzern & Denken*, die standardisierten Fragebogendaten und ihre quantitative Auswertung allein nicht hätten aufdecken können. Der direkte Austausch ermöglichte dem Projektteam einen neuen Blick auf das Projekt und seine Zielgruppen und generierte Erkenntnisse für Wissenschaftsvarietés und ähnliche Formate. Darüber hinaus hatten sie auch Folgen für das weitere Evaluationsdesign im Projekt: Es wurde verstärkt auf qualitative Befragungen in einem Mixed-Methods-Ansatz gesetzt.

Appendix

Tab. A.1 Kodierschema für die Einordnung von Assoziationen zu Wissenschaft und Kunst

Kodierschema für Wissenschaftsassoziationen nach Ziegler et al. (2018)		**Kodierschema für Kunstassoziationen** (abgeleitet aus dem Kodierschema für Wissenschaftsassoziationen)
Kategorie	**Aspekte**	**Kategorie**
Systematik und Regelgeleitetheit	Umfassende Analysen Hypothesen, Theorien, Methoden Experimente Objektivität Beweisführung	Aussagen, die den künstlerischen Prozess betreffen
Ergebnis- und Erkenntnisorientierung	Entdeckungen Erfindungen Innovationen Erkenntnisse Lösungen von Problemen	Aussagen, die die Funktion von Kunst betreffen
Nachvollziehbarkeit und kollaboratives Arbeiten	Transparenz Kommunikation der Ergebnisse und Vorgehensweise an Öffentlichkeit und Kollegen Reproduzierbarkeit	Aussagen, die die Vermittlung von Kunst und ihre Rezeption durch Außenstehende betreffen
Kritisch-reflexive Dimension, Neutralität, Gemeinwohlorientierung	Hinterfragen und Überprüfen bestehender Aussagen Kritische Reflexion der eigenen Arbeit Unabhängigkeit von Dritten Verantwortung gegenüber der Gesellschaft	Aussagen, die die Rolle von Kunst in der Gesellschaft betreffen

Tab. A.2 Kodierschema für die Einordnung von Assoziationen zu Wissenschaftler:innen und Künstler:innen

Kodierschema für Assoziationen zu Wissenschaftler:innen nach Ziegler et al. (2018)		**Kodierschema für Assoziationen zu Künstler:innen** (abgeleitet aus dem Kodierschema für Assoziationen zu Wissenschaftler:innen)
Kategorie	**Aspekte**	**Kategorie**
Fähigkeiten	Wissen, Fachkompetenz, Erfahrung IQ, Talent, logisches Denken Formale Qualifikationen, (Aus-) Bildung Kreativität	Aussagen, die Fähigkeiten der Künstler:innen betreffen
Arbeitsweisen	Analytisch, objektiv Exakt Ergebnisoffen Ausdauernd, geduldig, diszipliniert	Aussagen, die Arbeitsweisen der Künstler:innen betreffen
Motive	Neugierde, Visionen Integrität, Ehrlichkeit, Ethik, Moral Interesse, Leidenschaft, Zielstrebigkeit, Ehrgeiz	Aussagen, die Motive der Künstler:innen betreffen
Soziale Eigenschaften	Teamplayer, Kommunikation Unabhängigkeit ggü. anderen Interessen, Gemeinwohlorientierung	Aussagen, die soziale Eigenschaften der Künstler:innen betreffen

Interview-Leitfäden

Leitfaden Vorher-Interview

Block 1: Erwartungen an Freizeit- und Kulturangebote

1. Wir alle gestalten unsere Freizeit sehr unterschiedlich und haben verschiedene Bedürfnisse und Ansprüche an diese Zeit. Wonach suchen Sie in Ihrer Freizeit?
2. Natürlich kann man seine Freizeit einerseits selbst gestalten, andererseits bieten sich einem auch viele Möglichkeiten in Form von Veranstaltungen und Kulturangeboten. Wenn Sie an die Zeit vor Corona zurückdenken: Nutzen Sie solche Angebote regelmäßig und gerne in Ihrer Freizeit oder eher nicht?
   – Wenn ja: Was für Angebote sind das? Wie häufig nehmen Sie diese wahr?

   – Nur bei Nachfrage, was damit gemeint ist: Museen, Theater, Lesungen, Vorträge, Stadtfeste, …
3. Wäre eine Varieté-Veranstaltung etwas, was Sie auf eigene Initiative besuchen würden oder eher nicht?
4. Stellen Sie sich vor, Sie planen, ein Wissenschaftsvarieté zu besuchen: Was würden Sie von so einem Programm erwarten?
   – Falls nur künstlerische Aspekte genannt: Und wie, denken Sie, unterscheidet sich die Veranstaltung von einem klassischen Varieté?
   – Falls nur wissenschaftliche Aspekte genannt: Und wie, denken Sie, unterscheidet sich die Veranstaltung von klassischen Wissenschaftsveranstaltungen?

Block 2: Vorstellungen von Wissenschaft und Kunst

1. Wenn Sie den Begriff „Wissenschaft" hören, welche Assoziationen kommen Ihnen da?
2. Welche Eigenschaften und Kompetenzen verbinden Sie mit einem Wissenschaftler oder einer Wissenschaftlerin?
3. Wenn Sie den Begriff „Kunst" hören, welche Assoziationen kommen Ihnen da?
4. Welche Eigenschaften und Kompetenzen verbinden Sie mit einem Künstler oder einer Künstlerin?

Leitfaden Nachher-Interview

Block 1: Vorstellungen von Wissenschaft und Kunst

1. Wenn Sie den Begriff „Wissenschaft" hören, welche Assoziationen kommen Ihnen da?
2. Welche Eigenschaften und Kompetenzen verbinden Sie mit einem Wissenschaftler oder einer Wissenschaftlerin?
3. Und wenn Sie den Begriff „Kunst" hören, welche Assoziationen kommen Ihnen da?
4. Welche Eigenschaften und Kompetenzen verbinden Sie mit einem Künstler oder einer Künstlerin?

Block 2: Bewertung des Programms

1. Ganz allgemein und aus dem Bauch heraus: Wie hat Ihnen die Veranstaltung insgesamt gefallen?

2. Gibt es Teile des Programms, die Ihnen positiv in Erinnerung geblieben sind?
   – Wenn ja: Welche und warum?
3. Gibt es Teile des Programms, die Ihnen negativ in Erinnerung geblieben sind?
   – Wenn ja: Welche und warum?
4. In unserem letzten Gespräch hatte ich Sie gefragt, was Sie sich unter einem Wissenschaftsvarieté vorstellen und was Sie sich von so einer Show erhoffen würden. Hat die Veranstaltung Ihre Erwartungen getroffen oder eher nicht?
   – Wenn ja: Inwieweit?
   – Wenn nein: Was hat Sie überrascht, inwieweit wurden Ihre Erwartungen nicht getroffen?
   – Falls keine Erwartungen genannt wurden im Vorgespräch: Würden Sie sagen, was Sie gesehen haben, das passt zum Begriff Wissenschaftsvarieté?
   – Wie hat Ihnen die Verbindung von Kunst und Wissenschaft gefallen? Würden Sie sagen, Sie sind sowohl in Sachen wissenschaftliche Inhalte als auch künstlerische Darbietungen „auf ihre Kosten" gekommen oder eher nicht?
5. Wie haben Ihnen die wissenschaftlichen Erläuterungen zu den Mollusken gefallen (Vorstellung der Muscheln, Schnecken und Kopffüßer)?
6. Wie haben Ihnen die Anekdoten gefallen, die zu den verschiedenen Molluskenarten erzählt wurden?
7. Wie hat Ihnen die Einbindung der Publikumsfragen gefallen?
8. Wie haben Ihnen die musikalischen/künstlerischen Beiträge gefallen?
9. Haben Sie auch mal vorgespult oder sich anderen Dingen zugewendet?
   – Wenn ja: Können Sie sich erinnern, was zu dem Zeitpunkt im Programm passierte?

Block 3: Künftige Teilnahme und Verbesserungsvorschläge

1. Würden Sie an künftigen Veranstaltungen wieder teilnehmen (online oder real) oder eher nicht?
2. Haben Sie Wünsche an zukünftige *Glitzern & Denken*-Veranstaltungen oder Ideen, wie man das Programm für Sie noch attraktiver gestalten könnte?
   – …was die Inhalte betrifft?
   – …was den Ablauf/die Struktur des Programms betrifft?
   – …was die technische Umsetzung betrifft?

Literatur

Döring N, Bortz J (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, 5. Aufl. Springer, Berlin

Ziegler R, Kremer B, Weißkopf M (2018) Medizin und neue Technologien, Analysen und Erkenntnisse, Intelligenz und Ausdauer – Welche Vorstellung hat die Bevölkerung von Wissenschaft und Forschenden? Ergebnisse der offenen Fragestellungen im Wissenschaftsbarometer 2017. Wissenschaft im Dialog gGmbH, Berlin

Imke Hedder arbeitet für die deutsche Organisation für Wissenschaftskommunikation Wissenschaft im Dialog (WiD) im Bereich Qualität & Transfer. Als Teil der Impact Unit führte sie Analysen zur Evaluationspraxis in der deutschen Wissenschaftskommunikation durch und entwickelte Evaluationstools für Praktiker:innen. Inzwischen ist sie für die Evaluation verschiedener WiD-Projekte zuständig.

Ricarda Ziegler ist Leiterin des Bereichs Qualität & Transfer bei Wissenschaft im Dialog (WiD) – der deutschen Organisation für Wissenschaftskommunikation. Sie verantwortet dort u. a. die Impact Unit, die sich Fragen der Wirkung und Evaluation von Wissenschaftskommunikation widmet. Außerdem leitet sie das bevölkerungsrepräsentative Wissenschaftssurvey Wissenschaftsbarometer. Ricarda Ziegler hat einen Hintergrund in der Politikwissenschaft.

Bonnie Dietermann ist Programm-Managerin in der Abteilung Ausstellung am Museum für Naturkunde Berlin. Sie befasst sich mit partizipativen sowie co-kreativen Formaten an der Schnittstelle von Wissenschaft, Gesellschaft und Kunst. Bei dem Wissenschaftsvarieté Glitzern & Denken war sie für das Projektmanagement zuständig.

David Ziegler leitet das künstlerisch-wissenschaftliche Projekt Glitzern & Denken am Museum für Naturkunde Berlin. Sein Interesse gilt Aktivitäten an den Schnittstellen von Wissenschaft, Kunst, Gesellschaft und Politik.





Beobachtungen in der Evaluation von Wissenschaftskommunikation

Grundlagen und Praxis

André Weiß

Zusammenfassung

Der vorliegende Beitrag zeigt grundlegende Kriterien der sozialwissenschaftlichen Methode der Beobachtung auf und ordnet ihr Potenzial im Rahmen der Evaluation von Wissenschaftskommunikation ein. Der Beitrag geht dabei hauptsächlich auf eventbezogene Formate der externen Wissenschaftskommunikation ein, nennt aber auch Potenziale der (teilnehmenden) Beobachtung bei der Evaluation der Entstehung journalistischer Formate oder im Rahmen interner Wissenschaftskommunikation. Es wird insbesondere zwischen qualitativ und quantitativ ausgeprägten Formen der Beobachtung unterschieden und dargelegt, welche Schritte für eine erfolgreiche Durchführung zu beachten sind. Dabei werden auch Probleme und Herausforderungen bei der Anwendung der Methode genannt, derer durch die Beachtung bestimmter Gütekriterien vorgebeugt werden kann. Abschließend nennt der Beitrag Anwendungsszenarios und Beispiele für Beobachtungen im Rahmen von evaluativer Wissenschaftskommunikationsforschung.

Die (teilnehmende) Beobachtung gehört bereits seit langer Zeit zum gängigen Methodenrepertoire der empirischen Sozialwissenschaften. So können die Anfänge etwa in Paul Lazarsfelds Studie *Die Arbeitslosen von Marienthal* aus den 1930er Jahren gesehen werden, in der mithilfe von Beobachtungen erstmals in größerem Ausmaß systematisch quantitative Daten, wie die Gehgeschwindig-

A. Weiß (✉)
Department für Wissenschaftskommunikation, Institut für Technikzukünfte (ITZ), Karlsruher Institut für Technologie (KIT), Karlsruhe, Deutschland
E-Mail: andre.weiss@kit.edu

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,* https://doi.org/10.1007/978-3-658-39582-7_9

keit der Anwohner:innen, aber insbesondere auch qualitative Daten, etwa über die von Arbeitslosen selbst organisierten Veranstaltungen, erhoben und ausgewertet wurden (Jahoda et al. 2021; vgl auch Gehrau 2017; Heiser 2018).

Inzwischen wurde die Methode in zahlreichen weiteren Varianten beschrieben und nimmt zuletzt auch bei der Untersuchung von Wissenschaftskommunikation einen bedeutenderen Platz ein. Versteht man Wissenschaftskommunikation, wie die Herausgeber:innen des vorliegenden Bandes, nach Schäfer, Kristiansen und Bonfadelli als „alle Formen, von auf wissenschaftliches Wissen oder wissenschaftliche Arbeit fokussierter Kommunikation, sowohl innerhalb als auch außerhalb der institutionalisierten Wissenschaft, inklusive ihrer Produktion, Inhalte, Nutzung und Wirkungen“ (2015), so ergeben sich zahlreiche Szenarien, in denen die Beobachtung oft als begleitende, aber auch als zentrale Erhebungsmethode angewandt werden kann. Sei es im Rahmen von Dialog- oder Partizipationsformaten, wie Bürger:innenkonferenzen, bei Ausstellungen in Museen bzw. Science Centern (Fähnrich 2017), aber auch bei klassischen Formaten der strategischen Wissenschaftskommunikation, wie Pressekonferenzen oder Ähnlichem. Immer dort, wo Wissenschaftskommunikation als Event oder genuines Ereignis stattfindet, liegt die Form der Beobachtung als mögliche Evaluationsmethode auf der Hand, da es in einem zeitlich und lokal begrenzten Rahmen augenscheinlich *etwas zu beobachten* gibt.

Beobachtung kann jedoch aus verschiedenen Fachkulturen ganz unterschiedlich angeleitet werden und je nach Erhebungsanlage auch zu variierenden Ergebnissen kommen. Dieser Beitrag gibt daher einen Überblick über verschiedene Formen der Beobachtung und zeigt auf, wie diese bei der Evaluation von Wissenschaftskommunikation systematisch eingesetzt werden können, um eine Vergleichbarkeit und Bewertung zu ermöglichen und Wissenschaftskommunikation nachhaltig zu verbessern. Der Beitrag geht dabei hauptsächlich auf eventbezogene Formate der externen Wissenschaftskommunikation als wohl geläufigstem Anwendungsgebiet ein, nennt aber auch Potenziale der (teilnehmenden) Beobachtung bei der Evaluation der Entstehung journalistischer Formate oder auch im Rahmen interner Wissenschaftskommunikation.

1 Warum eignen sich Formen der wissenschaftlichen Beobachtung für die Evaluation von Wissenschaftskommunikation?

Bei der sozial- und kommunikationswissenschaftlichen Untersuchung von Wissenschaftskommunikation ist zuletzt vermehrt eine an der soziologischen Handlungstheorie orientierte Haltung eingenommen worden, das zeigen etwa

Beiträge wie der von Heinz Bonfadelli (2017) im Band *Forschungsfeld Wissenschaftskommunikation.* Demzufolge sind „theoretische Auseinandersetzung mit dem *Verhältnis von Wissenschaft und Laien,* aber auch die empirische Forschung dazu […] durch unterliegende meta-theoretische Konzepte zum Verhältnis von Wissenschaft, Medien und Laien beeinflusst, und zwar bezüglich *Fragestellungen* und *Publikumskonzeptionen,* aber nicht zuletzt auch bezüglich *normativer* Bewertungen." (Bonfadelli 2017, S. 84). Insbesondere in der Konzeption von „Wissenschaft auf dem Marktplatz" werden Bürger:innen Bonfadelli zufolge als „quasi gleichberechtigte Partner" (2017, S. 86) angesehen und Kommunikationssituationen werden vermehrt als zweiseitig asymmetrisch aufgefasst. Um die aktive Informationssuche und Wissensaneignung durch Laien, und insbesondere deren Einstellungsbildung zu untersuchen, werden Bonfadelli zufolge komplexere Modelle der Wirkung[1] erforderlich, weshalb auf der Mikroebene das kommunikative, soziale, sowie medienbezogene Handeln einzelner Akteur:innen in den Blick der Wissenschaftskommunikationsforschung gerät (Bonfadelli 2017).

Ebendieses Handeln ist auch der Gegenstand der Erhebungsmethode Beobachtung (Lamnek und Krell 2016). Das individuelle Verhalten steht für sie im Vordergrund. Gegenüber der Befragung, als meistverbreitete Erhebungsform bei der Evaluation von Wissenschaftskommunikation (Ziegler et al. 2021), die in der Regel nach bzw. vor und nach einem darin thematisierten Ereignis stattfindet, ermöglicht die Beobachtung eine mit dem zu untersuchenden Ereignis zeitgleiche Erfassung von sozialem Verhalten (Pürer 2009).

Bei der Evaluation von geplanten Formaten der Wissenschaftskommunikation ermöglicht der Einsatz von Beobachtung der beteiligten Akteur:innen insbesondere, Widersprüche und Unstimmigkeiten im Konzept, aber auch Potenziale unmittelbar im Moment der Kommunikation zu identifizieren, was beim Einsatz anderer, meist retrospektiver Erhebungsmethoden verwehrt bleiben würde (Pellegrini 2021, S. 311). Es verwundert daher nicht, dass die Beobachtung bereits vielfach bei der Evaluation von Wissenschaftskommunikation zum Einsatz kommt. Allzu oft wird sie jedoch als wenig reflektierte ergänzende Methode genutzt, die beiläufig erwähnt wird, vermeintlich wenig Ressourcen in Anspruch nehmen soll und nicht systematisch umgesetzt wird. Beispiel dafür sind unter anderen die explorative Untersuchung wissenschaftsorientierter Ausstellungshäuser von Freericks und Kolleg:innen (2018) oder auch das Konzept für eine

---

[1] Für verschiedene Wirkungsebenen der Wissenschafts- und Risikokommunikation vgl. Bonfadelli (2017, S. 88). Bonfadelli nennt hier etwa Mediennutzung und Rezeptionsprozesse, aber auch Wissen(sstand), Informiertheit, Einstellungen, Vertrauen und weitere.

prozessbegleitende Evaluation von Kommunikations- und Partizipationsformaten im Themenfeld Bioökonomie (Lux und Theiler 2021).

Bisher besteht das Problem, dass der eher willkürliche Einsatz von Beobachtung bei der Evaluation von Wissenschaftskommunikation einer Vergleichbarkeit verschiedener Projekte im Sinne einer nachhaltigen Verbesserung von Wissenschaftskommunikation im Weg steht (Ziegler et al. 2021). Der methodisch angeleitete Einsatz insbesondere standardisierter Beobachtungsverfahren kann eben diese Möglichkeit einer Vergleichbarkeit bieten. Andererseits vermögen aber nur offene, nicht-standardisierte Beobachtungsformen einen tieferen Einblick in komplexe Wirkungszusammenhänge, etwa bei der Wissensaneignung oder Einstellungsbildung, und individuelles Verhalten zu ermöglichen. Zu entscheiden, welche Form der Beobachtung angewandt werden soll, ist nicht immer einfach. Grundlage für die Entscheidung bildet die Fragestellung eines Evaluationsvorhabens, die auch den zentralen Unterschied zwischen alltäglicher Beobachtung und wissenschaftlicher Beobachtung ausmacht: Egal ob standardisiert oder nicht-standardisiert, wissenschaftliche Beobachtung ist stets auf eine Forschungsfrage ausgerichtet (Heiser 2018) und geht damit fokussiert vor.

Im Folgenden werden zunächst allgemein verschiedene Formen der wissenschaftlichen Beobachtung vorgestellt, die der Untersuchung des Gegenstands Wissenschaftskommunikation zweckdienlich sind, bevor abschließend Beispiele der Anwendung aufgezeigt werden.

2 Welche Formen der Beobachtung gibt es, und wofür sind sie geeignet?

In den Kommunikationswissenschaften kommt die Beobachtung als empirische Erhebungsmethode insbesondere in der Kommunikatorforschung, der Mediennutzungsforschung und der Rezeptionsforschung zum Einsatz (Pürer 2009). So können Wissenschaftler:innen, Organisator:innen und Moderator:innen ebenso zentraler Gegenstand von Beobachtungen im Setting Wissenschaftskommunikation sein, wie Besucher:innen, Leser:innen, Zuschauende und Zuhörende. Schon anhand dieser Gliederung wird deutlich: Im Zentrum der Beobachtung stehen Akteur:innen und ihr situationsbedingtes Handeln. Angesichts der Breite der eingangs beschriebenen Wissenschaftskommunikationslandschaft ist im Rahmen einer Evaluation von Wissenschaftskommunikation aber auch eine Anwendung der Methode im Geiste anderer Sozialwissenschaften, etwa der Soziologie oder der empirischen Kulturwissenschaften denkbar. Unabhängig von der grundlegenden

Fachdisziplin und der Frage nach einem standardisierten oder offen angelegten Vorgehen, hat die Methode stets zum Ziel „Alltagsphänomene [zu] explorieren und [zu] beschreiben, um daraus wissenschaftliche, insbesondere theoretische Aussagen entwickeln oder die Gültigkeit entsprechender wissenschaftlicher Aussagen anhand von Alltagsphänomenen überprüfen zu können" (Gehrau 2017, S. 19). Dafür bedarf es immer eines systematischen und geplanten Vorgehens, das auf eine wissenschaftliche Fragestellung hin ausgerichtet ist.

Der Entscheidung für die Beobachtung als – möglicherweise auch nur ergänzende – Erhebungsmethode im Rahmen eines Evaluationsvorhabens geht also immer die Identifikation eines Erkenntnisinteresses und eines damit verbundenen Evaluations-, sowie Erhebungsdesigns voraus. Es stellen sich vorab die wichtigen Fragen nach den Motiven der Evaluation (Dokumentation/Weiterentwicklung/Wissensgenerierung), der Ausrichtung der zentralen Fragestellung(en) (explorativ/explanativ) und auch nach der Verortung zwischen formativer (auf eine laufende Verbesserung ausgerichtete) und summativer (die Ergebnisse bilanzierende) Evaluation (Impact Unit und Wissenschaft im Dialog 2021a). Je nach Ausrichtung des Evaluierungsverfahrens ist anschließend auch die Ausgestaltung der Beobachtung zu modifizieren, und diese darüber hinaus auch in Einklang mit anderen angewandten Erhebungsmethoden zu bringen. Denn die Beobachtung kommt zumeist im Rahmen der Methodentriangulation (siehe auch Gabriel et al. in diesem Band) zum Einsatz und wird besonders häufig mit der Befragung kombiniert.

2.1 Klassifikationskriterien

Orientierung in der Planung einer Beobachtungsstudie bieten Klassifikationskriterien, die sich zum Großteil übereinstimmend in der breiten Beobachtungsliteratur wiederfinden (bspw. Lamnek und Krell 2016; Thierbach und Petschick 2019). Hier sind insbesondere sechs Klassifikationskriterien nach Döring et al. (2016) zu nennen, anhand derer richtungsweisende Fragen im Planungsprozess aufgestellt und bearbeitet werden können, weshalb sie hier auch im Einzelnen geschildert sind (Döring et al. 2016). Als Hilfestellung im Planungsprozess von Evaluationsverfahren innerhalb derer die Beobachtung gegebenenfalls zum Einsatz kommen soll, empfiehlt es sich außerdem die Handreichungen zu Rate zu ziehen, die zuletzt im Rahmen des Projekts Impact Unit entstanden sind (Impact Unit und Wissenschaft im Dialog 2021b).

- Eine grundlegende Entscheidung ist in Bezug auf den *Strukturierungsgrad der Beobachtung* zu treffen. Der (voll)strukturierten Beobachtung, bei der mithilfe eines standardisierten Beobachtungsbogens nur vorher klar definierte Variablen erhoben werden, steht die teil- oder unstrukturierte Beobachtung gegenüber, in der mit wenigen oder gänzlich ohne Beobachtungsrichtlinien beobachtet wird. In der Praxis wird der Strukturierungsgrad auch durch die Unterscheidung zwischen quantitativer und qualitativer Beobachtung zum Ausdruck gebracht.
- *Gegenstand der Beobachtung* sind in der Regel Merkmale und Verhaltensweisen anderer Personen oder auch der Beobachtenden selbst. Daher wird zwischen Fremdbeobachtung und Selbstbeobachtung unterschieden. Erstere ist bei der Evaluation von Wissenschaftskommunikation die geläufigere Form, da das Forschungsinteresse zumeist auf beteiligten Akteur:innen von verschiedenen Formaten ruht. Denkbar ist aber auch, etwa im Rahmen formativer Evaluationen von Forschungsprojekten, eine Selbstbeobachtung von Forschenden, die in der Regel qualitativ als Autoethnographie durchgeführt wird. Diese spielt aber eine eher untergeordnete Rolle.
- Auch lässt sich unterscheiden nach der *Direktheit der Beobachtung*. In der Regel wird direkt beobachtet, also mit Kontakt zwischen Beobachtenden und Beobachteten. Indirekte Beobachtung kommt ohne einen solchen aus, da nicht das Verhalten selbst, sondern Verhaltensspuren beobachtet werden. Im Setting Wissenschaftskommunikation können diese beispielsweise in Form von mehr oder weniger stark abgenutzten Ausstellungsobjekten in Hands-On-Ausstellungen oder der Inanspruchnahme eines Buffets nach (!) einer Bürger:innenkonferenz bestehen.[2]
- Bezüglich des *Ortes der Beobachtung* wird mit Blick auf das natürliche Lebensumfeld auf der einen und der kontrollierten Laborsituation auf der anderen Seite in Feld- und Laborbeobachtungen unterschieden. Relevant kann außerdem die Unterscheidung zwischen Offline- und Onlinebeobachtung sein. In jedem Fall wird die Beobachtung mit diesem Kriterium lokal begrenzt, da sie andernfalls nicht fokussiert erfolgen könnte.
- Ein wichtiges Unterscheidungsmerkmal von Beobachtungsformen ist außerdem der *Involviertheitsgrad der Beobachtung*. Im Rahmen einer Fremd-

[2] Stellenweise auch in der Beobachtungsliteratur beschrieben und der Methode der Beobachtung zugeordnet sind Verfahren, mit denen Verhaltensspuren in der Nutzung digitaler Medien nachverfolgt werden können, wie etwa die Nutzung von Diensten wie Twitter oder ähnlichem. Gehrau widmet dieser innovativen Umsetzung der Methode Beobachtung mehrere Kapitel (Gehrau 2017, S. 153 ff.). Döring et al. (2016) dagegen ordnen diese Form explizit der Dokumentenanalyse zu. Siehe auch Hempel sowie Bruckermann und Greving in diesem Band.

beobachtung können Forschende nicht-teilnehmend beobachten oder teilnehmend beobachten. Sofern sie teilnehmen kann diese Teilnahme darüber hinaus aktiv (primär teilnehmend) oder passiv (primär beobachtend) erfolgen.
- Zuletzt nennen Döring et al. (2016) die *Transparenz der Beobachtung*. Mit diesem Kriterium sind insbesondere forschungsethische Fragen verbunden, sofern die Beobachtung nicht offen (mit dem Wissen der Beobachteten stattfindet), sondern verdeckt durchgeführt wird.

Nach Gehrau (2017) gibt es neben diesen zentralen Kriterien noch weitere forschungspraktische Überlegungen, die es anzustellen gilt: Beobachten Forschende selbst, oder werden externe Beobachtende beauftragt? Soll die Beobachtungen mit einem Stimulus erfolgen? Findet die Beobachtung unvermittelt, also in Anwesenheit von Beobachtenden und Beobachteten, oder vermittelt statt, etwa in Form einer Aufzeichnung? Erfolgt die Protokollierung der Beobachtung manuell oder apparativ-automatisiert?[3]

Zweckdienlich miteinander kombinieren lassen sich nicht alle Ausprägungen der oben genannten Kriterien. Stattdessen haben sich bestimmte Kombinationen etabliert. So identifiziert Gehrau (2017) im Rahmen einer Metastudie etwa das Setting einer Fremdbeobachtung, die von Forschenden selbst durchgeführt wird, welche dabei jedoch nicht teilnehmen, als Grundtypus von Beobachtungen im Rahmen kommunikationswissenschaftlicher Forschung, bei dem zudem häufig unvermittelt, strukturiert, indirekt und wissentlich beobachtet wird. Die Beobachtung interpersonaler Kommunikation findet dagegen häufig in Form von Feldstudien statt, in denen „offen, wissentlich, direkt, unvermittelt, ohne Stimulus und nie apparativ" (Gehrau 2017, S. 139 f.) beobachtet wird. Disziplinenübergreifend lässt sich diese grundsätzliche Herangehensweise immer wieder auffinden. Erste Orientierung bietet damit insbesondere der Grad der Strukturierung einer Beobachtung: Für das eine Evaluationsdesign eignen sich eher qualitative, für ein anderes eher quantitative Formen der Beobachtung, innerhalb derer die oben genannten Kriterien dann entsprechend auszustaffieren sind. Diese beiden Ausrichtungen werden im Folgenden genauer gegenübergestellt.

[3] Gehrau (2017) zählt etwa auch technische Verfahren wie das „Eyetracking" zur Methode der Beobachtung in der Kommunikations- und Medienwissenschaft, was in dem vorliegenden Beitrag jedoch nicht weiter behandelt wird (vgl. dazu Niemann und Scheuermann in diesem Band).

2.2 Quantitative Beobachtung

Die vollstrukturierte quantitative Beobachtung folgt einem linearen Vorgehen im Forschungsprozess. Demnach werden in einem vorab festgelegten Untersuchungsfeld, das zeitlich und räumlich eingegrenzt wird, numerische Daten erfasst, die anschließend statistisch ausgewertet werden (Thierbach und Petschick 2019, S. 1169). Eine quantitativ ausgerichtete Beobachtung eignet sich daher in der Regel insbesondere dann, wenn zuvor klar definierte (Teil-)Ziele eines Projekts oder einer Maßnahme gemessen und überprüft oder Hypothesen getestet werden sollen, die bereits auf einem vergleichsweise breiten Vorwissen basieren. So kommt diese Form der Beobachtung eher in Evaluationsdesigns zum Einsatz, in denen Evaluationsfragen explanativ überprüft werden sollen.

Dabei kann die Beobachtung im Komplexitätsgrad jedoch variieren. Neben einfachen Aspekten des beobachteten Vorgangs, wie „Häufigkeit, Dauer und/oder Intensität" können durch „geprüfte und etablierte Instrumente" (Döring et al. 2016, S. 343 ff.) auch komplexere Verhaltensvorgänge gemessen werden. Diese häufig aus der Psychologie stammenden Beobachtungsinstrumente sind jedoch auch in ihrer Handhabung komplex und erfordern gegebenenfalls eine lange Einarbeitungszeit. Für die Evaluation von Wissenschaftskommunikation empfiehlt sich die Anwendung daher wohl nur in Zusammenarbeit mit erfahrenen Sozialwissenschaftler:innen oder Psycholog:innen. Eher bieten sich einfachere quantitative Beobachtungen an, mit denen aber bereits wesentliche Daten erhoben werden können, wie etwa quantifizierbare Merkmale der Beteiligung in partizipativen Bürgerformaten, des Besucher:innenverhaltens in Ausstellungen oder anderen Veranstaltungen mit Publikum. Dabei kann es sich beispielsweise um die Häufigkeit von Wortmeldungen, die Rede- oder Verweildauer handeln. Zuletzt kann auch eine nonreaktive Beobachtung von Verhaltensspuren quantitativ erfolgen, die im Sinne einer indirekten Beobachtung (siehe oben) ohne Kontakt zwischen Beobachtenden und Beobachteten auskommt.

Das wichtigste Erhebungsinstrument der quantitativen Beobachtung stellt der Beobachtungsbogen dar, der ausgerichtet auf Fragestellung und Beobachtungsgegenstand in hohem Maße strukturiert ist. Der Vorbereitung von Beobachtungsbögen geht die Feldbestimmung und Einteilung des Beobachtungsgegenstands in Beobachtungseinheiten (einzelne Aspekte, die beobachtet werden sollen) voraus. Anschließend folgt eine Operationalisierung des Beobachtungsbogens analog zum standardisierten Fragebogen (Thierbach und Petschick 2019, S. 1176 f.). Protokollbögen für eine standardisierte Beobachtung sollten dabei zwar alle notwendigen Kriterien erfüllen und Inhalte abdecken, gleichzeitig ist darauf zu

achten, dass insbesondere bei externer Beobachtung die Durchführung nicht durch einen zu wissenschaftlichen Sprachduktus und außerordentlich komplexe Protokollbögen erschwert wird. Generell sollte darauf geachtet werden, dass Beobachtungsbögen praktisch handhabbar sind und die Beobachtungssituation nicht behindern (Gehrau 2017, S. 51 ff.). Es empfiehlt sich eine Aufbereitung, die den Bogen „leicht visuell erfassbar" (Thierbach und Petschick 2019, S. 1176) macht und zu beobachtende Ereignisse anschaulich mit Beispielen verdeutlicht.

Dabei haben Beobachtungsbögen aber analog zum standardisierten Fragebogen Gütekriterien zu erfüllen. Im Rahmen einer ordentlichen Stichprobenziehung ist die Angemessenheit der Auswahl der Beobachtungeinheiten zu hinterfragen. Darüber hinaus sollte der Bogen die Objektivität der Beobachtung sichern: Zwei verschiedene Beobachtende sollten auf dessen Grundlage immer zum selben Ergebnis kommen. Ist das gegeben, kann auch die Reliabilität oder Messgenauigkeit der Beobachtung überprüft werden. Sie gibt an, „wie gut zwei Beobachtungen desselben Geschehens übereinstimmen" (Gehrau 2017, S. 54). Es ist also ratsam, bei quantitativen Beobachtungen immer mindestens zwei Beobachtende parallel beobachten zu lassen. Zuletzt gibt die Validität an, inwiefern das was zu beobachten geplant war auch tatsächlich beobachtet werden konnte. Interne Validität steht demnach für die Frage danach, ob Beobachtungsituation und Fragestellung kompatibel sind. Externe Validität dagegen ist dann gegeben, wenn sich die Ergebnisse verallgemeinern lassen und tatsächlich das im Interesse stehende Alltagshandeln beobachtet werden konnte (Döring et al. 2016; Gehrau 2017; Thierbach und Petschick 2019).

2.3 Qualitative Beobachtung

Die qualitative Beobachtung dagegen folgt einer iterativ-zyklischen Vorgehensweise und damit einer ganz anderen Forschungslogik. Ziel ist es nicht bestehende Hypothesen zu testen, sondern neue zu generieren. Die dabei erfassten Daten sind in der Regel verbaler Art und schlagen sich in einem Beobachtungsprotokoll nieder, das mithilfe von Auswertungsmethoden, wie der qualitativen Inhaltsanalyse, analysiert wird. Anhand der Ergebnisse kann dann für weitere Beobachtungssituationen das Untersuchungsfeld ausgeweitet oder eingegrenzt werden (Thierbach und Petschick 2019; siehe auch Döring et al. 2016). Die qualitative Beobachtung eignet sich damit insbesondere für Evaluationsdesigns, in denen Abläufe in Projekten explorativ erkundet und damit neues Wissen generiert werden soll.

Zwei für die Evaluation von Wissenschaftskommunikation relevante Formen der qualitativen Beobachtung sind Beobachtungen mit geringem Komplexitätsgrad sowie ethnographische Feldforschungen. Erstere kommen dann zum Einsatz, wenn nur Einzelaspekte des beobachteten Feldes im Forschungsinteresse stehen. Das Vorgehen ist wegen der Fokussierung teil-strukturiert, beschreibt das zu beobachtende Forschungsproblem dabei aber „in seinen qualitativen Merkmalen" (Döring et al. 2016, S. 334). Ethnographische Feldforschung nimmt sich dagegen eines ganzen zu beobachtenden Feldes an, wird in der Regel über einen längeren Zeitraum durchgeführt und geht dabei zunächst wenig bis gar nicht fokussiert vor. Eine Fokussierung findet oft erst nach ersten Auswertungsdurchgängen, beispielsweise im Sinne einer qualitativen Inhaltsanalyse, statt. Auch wenn qualitative Forschung vorab keine Kategorien aufstellt, so ist das Verfahren trotzdem an einem theoretischen Interesse ausgerichtet, weshalb auch hier Beobachtungseinheiten definiert werden können. Anders als in der quantitativen Beobachtung handelt es sich dabei aber nicht um kleinste Verhaltenseinheiten, sondern meist um „soziale Situationen, die dann erst in der Analyse in einzelne Bestandteile zerlegt werden" (Lamnek und Krell 2016, S. 553). Die Beobachtungseinheiten qualitativer Beobachtungen sind komplexerer Natur, da zum Beispiel „Interaktionsmuster anstelle einzelner Verhaltensweisen" (Döring et al. 2016, S. 334) beobachtet werden.

Bei der qualitativen Beobachtung spielt das Austarieren des Involviertheitsgrades der Beobachtenden eine wichtige Rolle. Es wird zwischen primärer Beobachtung (passive Teilnahme) und primärer Teilnahme (aktive Teilnahme) unterschieden. In der Ethnographie wird die Teilnahme als Voraussetzung angesehen, um einen Forschungsgegenstand überhaupt erschließen zu können. Gehrau (2017) empfiehlt dagegen – und diese Empfehlung kann überwiegend auch für die Evaluation von Wissenschaftskommunikation gelten – die passive Teilnahme als geeigneten Mittelweg, um eine Alltagssituation weder durch die bloße Beobachtung und damit als Fremdkörper, noch durch aktive Handlungen im Feld zu verstellen (Gehrau 2017, S. 31 f.; siehe auch Döring et al. 2016).

Sowohl bei Beobachtungen mit geringem Komplexitätsgrad, als auch bei ethnografischen Feldbeobachtungen erfolgt die Protokollierung anders als im vollstrukturierten Vorgehen in eigenen Worten im Rahmen von Feldnotizen, die unmittelbar im Anschluss an die Beobachtung in ein Feldprotokoll überführt werden sollten. Neben der Beschreibung der Beobachtungen und ihrer Kontextualisierung fließen dabei auch methodische und theoretische Reflexionen ein (Döring et al. 2016). Aus diesem Grund kann die qualitative Beobachtung, insbesondere in Form der ethnographischen Feldbeobachtung, nicht oder nur sehr bedingt an externe Beobachter abgegeben werden. Abhängig vom Komplexitäts-

grad der Fragestellungen, die die Evaluation anleiten, ist eine interne Form, also die Durchführung der Beobachtung durch Forschende selbst, womöglich besser geeignet.

Zuletzt hat auch die qualitative Beobachtung bestimmten Gütekriterien zu genügen. Dabei muss die Beobachtung intersubjektiv nachvollziehbar, das heißt durch andere Forschende überprüfbar, sein. Dabei spielt das Beobachtungsprotokoll eine wichtige Rolle. Dieses genügt dann den Ansprüchen, wenn eine kontinuierliche Reflexion des Forschungsprozesses und der Beobachtenden selbst erfolgt ist (Thierbach und Petschick 2019). Gehrau (2017) führt außerdem die Möglichkeit der Kontrolle durch die Beobachteten selbst an, indem man Protokolle nachträglich gegenlesen lässt und dabei beispielsweise die Methode des Lauten Denkens anwendet.

2.4 Probleme und Herausforderungen bei Beobachtungen

Im Beobachtungsprozess können zahlreiche Probleme auftreten, die es je nach Beobachtungssetting zu adressieren gilt. Döring et al. (2016) zufolge können Fehler durch die Beobachtungssituation oder durch die beobachtende Person entstehen. So besteht ein allgemeines und weitreichendes Problem in der Reaktivität der Beobachtungssituation. Direkte Beobachtungen bergen die Gefahr einer nicht-intentionalen Beeinflussung des Verhaltens von Beobachteten, sei es auf deren Seite durch das Bewusstsein, dass sie beobachtet werden, oder sei es aufgrund der aktiven Teilnahme durch die Beobachtenden und den damit verbundenen direkten Eingriff in die beobachtete Situation. Eine zentrale Herausforderung besteht dabei insbesondere bei teilnehmender Beobachtung in der Wahrung der analytischen Distanz, bei gleichzeitigem Versuch teilzunehmen und „sozial und kulturell konform zu handeln“ (Heiser 2018, S. 71).

Beobachtungsfehler können insbesondere auch in der praktischen Durchführung entstehen. So kann ein zu einfacher oder auch zu komplexer Beobachtungsbogen die ausreichende Protokollierung verhindern. Um solchen Beobachtungsfehlern vorzubeugen, empfehlen sich neben der fortdauernden Reflexion der eigenen Beobachtungstätigkeit und deren Einfluss auf den Beobachtungsgegenstand insbesondere die Durchführung eines Pretests der Beobachtungsmaßnahme und bei externen Beobachtern darüber hinaus geeignete Schulungsmaßnahmen (Döring et al. 2016; siehe auch Gehrau 2017).

Eine weitere häufig genannte Hürde im Beobachtungsprozess, die im Rahmen von projektbegleitender Evaluation jedoch eher vernachlässigt werden kann, stellt

der Feldzugang dar. Zwar können sich auch hier zu beobachtende Personen der Beobachtung verweigern, in der Regel ist der Zugang zum Feld aber im Rahmen der Anbindung der Forschung an ein zu evaluierendes Projekt gewährt.

Ein anderes Problem dagegen, das insbesondere bei teilnehmender Beobachtung besteht, ist für eine gelingende Evaluation durchaus relevant und sollte beachtet werden: Beobachtungen können sehr umfangreiche Datenmengen produzieren, für deren Auswertung auch Kapazitäten vorhanden sein müssen. Die Gefahr ist groß, dass in prozessbegleitenden Beobachtungen sehr viele Daten gesammelt werden, diese dann aber nicht mehr ausgewertet werden, weil im zeitlich begrenzten Auswertungsprozess den Ergebnissen anderer Methoden, wie etwa Befragungen, vermeintlich mehr Bedeutung beigemessen wird. Je nach Projektdauer und -mitteln bieten sich bei größeren Projekten Abschlussarbeiten oder auch Dissertationsvorhaben an, um das Problem der großen Datenmengen einzelner Beobachtungsformen bewusst adressieren zu können und die Methode der Beobachtung nicht von vornerein aus forschungspragmatischen Gründen ausschließen oder in ihrem Umfang stark begrenzen zu müssen. Dieser Ansatz ist jedoch nur zu empfehlen, wenn der Evaluation neben den üblichen Nutzungsansprüchen für die Praxis vorrangig ein wissenschaftliches Erkenntnisinteresse zugrunde liegt.

3 Anwendungsszenarios in Wissenschaftskommunikationsforschung und -evaluation

Je nach Ausrichtung einer Evaluation bieten sich nun verschiedene Formen der Beobachtung an. Nur kurz erwähnt seien an dieser Stelle Möglichkeiten in der internen Wissenschaftskommunikation, dem Wissenschaftsjournalismus und der PR-Forschung. In allen drei Bereichen finden sich Beispiele in der Forschung, in denen zumeist die teilnehmende Beobachtung zum Einsatz kommt. So existieren seit den 1980er-Jahren, etwa von Karin Knorr-Cetina (1983), Bruno Latour (Latour und Woolgar 1986) und anderen zahlreiche ethnographische Beobachtungsstudien zu informeller Wissenschaftskommunikation in verschiedenen Fachkulturen (Lüthje 2017). In Forschungsarbeiten zum Wissenschaftsjournalismus gehört die Redaktionsbeobachtung zum etablierten Methodenrepertoire (z. B. Lublinski 2004). Und zuletzt liegen auch für die PR-Forschung Beispiele vor, in denen Arbeitsabläufe mithilfe von Beobachtungen genauer untersucht werden, wie etwa die Dissertation von Howard Nothhaft, in der Kommunikationsmanager:innen beobachtet wurden, um die alltägliche

Strukturierung ihrer Tätigkeit offenzulegen (Nothhaft 2011; siehe auch Wehmeier und Raaz 2016). In allen hier genannten Fällen sind teilnehmende Beobachtungen mit hohem Komplexitätsgrad durchgeführt worden, in denen Forschende selbst über lange Zeiträume beobachtet und protokolliert und eine große Datenmenge gesammelt haben. Diese Beispiele können bei der Planung einer Evaluation in einem der genannten Bereiche Orientierung bieten – je nach Ressourcen ist die Umsetzung aber sicherlich einzugrenzen.

Ein weiterer Bereich, in dem die Ergebungsmethode der Beobachtung einen Platz hat, sind pädagogischen Angebote, die an die Wissenschaftskommunikation angrenzen. Insbesondere zur Bildung für nachhaltige Entwicklung (Hiller 2017; Lechner 2018), über Kinderunis (Bergs-Winkels und Ludwig 2006; Kretschmer 2017a, b), aber auch zu Schüler:innenlaboren (Schmitt-Sody 2013) finden sich beispielhafte Anwendungen. Hier haben die Beobachtungen entweder im Sinne einer begleitenden Maßnahme einen eher untergeordneten Stellenwert oder werden im Rahmen von Dissertationen als eine zentrale Erhebungsmethode herausgestellt.

Hauptsächlich findet die Beobachtung aber in der Erforschung externer Wissenschaftskommunikation ihre Anwendung. In der Evaluation von Hochschulkommunikation etwa wird sie als Erhebungsmethode insbesondere auf der Messbereich-Ebene des Outcomes im Rahmen von Veranstaltungen angewandt (Raupp und Osterheider 2019). Auch über die Hochschulkommunikation hinaus ist das *Wissenschaftsevent* der wohl häufigste Gegenstand wissenschaftlicher Beobachtung in der Wissenschaftskommunikationsforschung und -evaluation. So liegt etwa zum populären Abendformat des Science Slams eine qualitative Studie vor, in der das empirische Material aus Interviews durch „fokussiert ethnographische Beobachtungen" (Hill 2020, S. 155) ergänzt wird. Ebenfalls ergänzend zu Leitfadeninterviews wurde das Format „Plötzlich Wissen", das in Kneipen stattfand, mithilfe einer qualitativen Beobachtung mit eher geringem Komplexitätsgrad untersucht (Bittner 2018). Die Anwendung der Beobachtung im Zusammenhang mit Wissenschaftsevents nimmt in den vergangenen Jahren stetig zu, und ein frühes Beispiel dafür findet sich in der Evaluation der Science Week Austria. Hier wurden in zweiköpfigen Beobachtungsteams mithilfe strukturierter Beobachtungsleitfäden die Daten aus einem breiten Methodenrepertoire ergänzt (Felt et al. 2001). Weitere Beispiele im Zusammenhang mit Wissenschaftsevents, die sich an ein breites Publikum wenden und bei denen mit oft geringem Komplexitätsgrad beobachtet wird, finden sich im Zusammenhang mit Museums- und Ausstellungsuntersuchungen (Decristofero et al. 2016; Haywood 2017; Heil und Jakobs 2017; Munsch 2017).

Bei der Untersuchung von Beteiligungsformaten wie Citizen Science, Bürger:innen-Delphis oder ähnlichem finden sich in der Anwendung der wissen-

schaftlichen Beobachtung sowohl strukturierte als auch unstrukturierte Verfahren, wobei die Methode in den beiden hier beschriebenen Fällen mit anderen Methoden, insbesondere der Befragung kombiniert wird. Goldschmidt und Kolleg:innen (2012) beobachteten *sieben* verschiedene Dialog- und Beteiligungsformate im Rahmen des Projekts „Wissenschaft debattieren!“. Aufgrund ihres Vorgehens anhand eines vorstrukturierten Beobachtungsbogens war es ihnen möglich, quantitative Daten, wie die Anzahl von Redebeiträgen systematisch zu erheben und untereinander zu vergleichen (Goldschmidt et al. 2012). Bei diesem Beispiel ermöglicht also das strukturierte Vorgehen insbesondere den Vergleich verschiedener Formate anhand gleichbleibender Kriterien.

Metten und Bornheim (2021) dagegen schildern ihre teilnehmende Beobachtung als Teil einer „mikro-ethnographische[n] Herangehensweise“ (S. 26) zur Untersuchung von Partizipationskultur anhand *eines* Realexperiments. Die Methode der Beobachtung ermöglichte es in diesem Fall insbesondere „kollaborative mediale Praktiken“ (Metten und Bornheim 2021, ebd.) zu erfassen, die einen wesentlichen Bestandteil der untersuchten responsiven Wissenschaftskommunikation ausmachen. Die unvermittelt durchgeführte Beobachtung wurde dafür in Beobachtungsprotokollen dokumentiert und zusätzlich per Video aufgezeichnet, wesentliche Passagen nachträglich transkribiert und ausgewertet (Metten und Bornheim 2021). Dabei ist nicht nur eine große Datenmenge entstanden, die eine Anwendung im Rahmen mehrerer verschiedener Formate, wie im oben genannten Beispiel, unrealistisch erscheinen lässt. Auch wäre es kaum vorstellbar, solche Daten ohne eine vorhergehende Strukturierung über sieben verschiedene Formate hinweg sinnvoll auszuwerten und dabei den genannten Gütekriterien gerecht zu werden.

4 Hilfreiche Beispiele für Beobachtungsbögen

Bis hierher hat sich gezeigt, welches Potenzial die Methode der Beobachtung für die Evaluation von Wissenschaftskommunikation birgt, welche Herausforderungen sich in der Umsetzung stellen können und welche Anwendungsszenarien bisher bekannt sind. Ergänzend findet sich nun abschließend der Verweis auf zwei beispielhafte Beobachtungsbögen, die als Vorlage für eigene Evaluationsvorhaben dienen können. Wie oben beschrieben, spielt der Beobachtungsbogen insbesondere in quantitativen Verfahren, aber auch in teil-standardisierten qualitativen Beobachtungen eine zentrale Rolle. Aus diesem Grund finden sich abschließend zwei Beispiele für die Erarbeitung und Anwendung eines Beobachtungsbogens. Da das Beobachtungsprotokoll, das im

Rahmen einer teilnehmenden Beobachtung entsteht, dagegen zunächst möglichst offen sein sollte, wird hierfür kein explizites Beispiel angeführt.

Das erste Beispiel für einen Beobachtungsbogen stammt von Léonie Rennie und Terry McClafferty (1996) und ist dem *Handbook for Formative Evaluation of Interactive Exhibitions* entnommen, das eine Handreichung für die Evaluation interaktiver Ausstellungen bietet. In einem multimethodischen Forschungsdesign stellt die Beobachtung den ersten Schritt der Erhebung dar und dient der weiteren Fokussierung des Evaluationsvorhabens. Mithilfe eines standardisierten Beobachtungsbogens wird das Verhalten von Ausstellungsbesucher:innen beobachtet und dokumentiert, ob und wie sie mit der Ausstellung interagieren. Neben Geschlecht und Altersspanne wird die Involviertheit von Besucher:innen in vier Leveln („ignore", „attend", „engage", „use successfully") erhoben, sowie deren Interaktionen mit anderen Besucher:innen. Für eine einheitliche Dokumentation erläutert der Leitfaden die Level in einem Glossar und macht zudem Vorgaben, wie die Interaktionen auf dem Beobachtungsbogen darzustellen sind (Rennie und McClafferty 1996, S. 10 ff.). Eine Anwendung dieser Vorlage findet sich in der Evaluation einer Mitmachausstellung im Technischen Museum Wien durch Stephanie Trauner, die sich in einer Beobachtungsstudie mit Interaktionen zwischen Erwachsenen und Kindern im Kontext von Ausstellungen mit Hands-On-Elementen befasst hat (Trauner 2016).

Ein zweites Beispiel liefern Mitglieder der nordamerikanischen Association of Zoos & Aquariums (AZA) mit einem für „Measuring Empathy: A Collaborative Assessment Project" entwickelten standardisierten Beobachtungsbogen (Jackson und Khalil 2019). Dieser lässt sich in verschiedene Settings übertragen, ist dabei aber grundsätzlich auf die Frage nach Ausdrucksformen von Empathie und damit verbundenen Gefühlen gegenüber Tieren ausgerichtet. Nichtsdestotrotz kann auch er herangezogen werden, um die Eigenschaften eines Beobachtungsbogens zu verdeutlichen und als Vorlage für eigene Vorhaben zu dienen. Neben einer einleitenden Einordnung des Beobachtungsbogens nach Einsatzgebiet, Ressourcenbedarf und zentralen Fragestellungen, bietet die Vorlage insbesondere einen Bewertungsrahmen, in dem empathische und emotionale Ausdrücke – die mithilfe der Methode beobachtet werden sollen – in sieben verschiedene Kategorien aufgeteilt werden, denen jeweils Indikatoren für die Beobachtung zugeordnet werden. Der Bewertungsrahmen dient zudem als Codebuch für einen in der Durchführung leicht zu handhabenden Beobachtungsbogen, in dem sich die Kategorien wiederfinden.

Die beiden hier genannten Beispiele können Orientierung bei der Erarbeitung einer standardisierten oder halbstandardisierten Beobachtung und einem dafür anzuwendenden Beobachtungsbogen im Rahmen einer Evaluation von Wissen-

schaftskommunikation bieten. Sofern die Methode der Beobachtung also basierend auf den Evaluationszielen an den unter Abschn. 2.1 genannten Kriterien ausgerichtet wird, kann sie ein Evaluationsdesign im Sinne der Methodentriangulation komplettieren.

Literatur

Bergs-Winkels D, Ludwig S (2006) Die Uni in der Kinder-Uni: Eine Begleitstudie zur Münsteraner Kinder-Uni. LIT, Münster

Bittner L (2018) Guerilla-Wissenschaftskommunikation in der Kneipe: Untersuchung am Projekt „Plötzlich Wissen!“, Science In Presentations/Arbeitsberichte, Bd 4. Karlsruher Institut für Technologie (KIT). https://doi.org/10.5445/IR/1000132272

Bonfadelli H (2017) Handlungstheoretische Perspektiven auf die Wissenschaftskommunikation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer VS, Wiesbaden, S 83–105. https://doi.org/10.1007/978-3-658-12898-2_5

Decristofero B, Hopmann S, Kartschnig T, Seebauer L, Swertz C, Decristoforo B, Katschnig T (Hrsg) (2016) Hands-On im Technischen Museum Wien: Konzeption und Evaluierung der Mitmachausstellung „In Bewegung“, 1. Aufl. nap, new academic press, Wien

Döring N, Bortz J, Pöschl-Günther S (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer, Berlin

Fähnrich B (2017) Wissenschaftsevents zwischen Popularisierung, Engagement und Partizipation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer VS, Wiesbaden, S 165–182. https://doi.org/10.1007/978-3-658-12898-2_9

Felt U, Müller A, Schober S (2001) Evaluierung der Science Week @ Austria 2001. https://repository.fteval.at/290/Gesehen. Zugegriffen: 28. Juni 2022

Freericks R, Brinkmann D, Theile H (2018) Wissenswelten 3.0: Eine explorative Untersuchung von Entwicklungsmöglichkeiten im Bereich der wissenschaftsorientierten Ausstellungs- und Bildungshäuser – mit besonderem Fokus auf Trends der Digitalisierung und einem Wandel des Lernverhaltens, 1. Aufl. Institut für Freizeitwissenschaft und Kulturarbeit, Bremen

Gehrau V (2017) Die Beobachtung als Methode in der Kommunikations- und Medienwissenschaft, 2. Aufl. utb GmbH, Stuttgart. https://doi.org/10.36198/9783838548418

Goldschmidt R, Scheel O, Renn O (2012) Zur Wirkung und Effektivität von Dialog- und Beteiligungsformaten [WorkingPaper]. http://elib.uni-stuttgart.de/handle/11682/5569. Zugegriffen: 28. Juni 2022

Haywood N (2017) The building bridges research project at the London science museum: using an ethnographic approach with under-represented visitor groups. Archaeol Int 20:69–73. https://doi.org/10.5334/ai.356

Heil A, Jakobs EM (2017) DC-Technologien für jedermann?: Eine empirische Studie zur Wissensvermittlung in Ausstellungen: Projektbericht. RWTH Aachen University

Heiser P (2018) Die Arbeitslosen von Marienthal. Oder: Die Anfänge qualitativer Sozialforschung. In: Heiser P (Hrsg) Meilensteine der qualitativen Sozialforschung: Eine Einführung entlang klassischer Studien. Springer VS, Wiesbaden, S 53–90. https://doi.org/10.1007/978-3-658-18557-2_2

Hill M (2020) Wissenschaft und Öffentlichkeit im Zeichen der Digitalisierung. In: Niemann P, Bittner L, Hauser C, Schrögel P (Hrsg) Science-Slam: Multidisziplinäre Perspektiven auf eine populäre Form der Wissenschaftskommunikation. Springer VS, Wiesbaden, S 149–180. https://doi.org/10.1007/978-3-658-28861-7_9

Hiller S (2017) Nachhaltigkeit lernen II – Kinder gestalten Zukunft. Ergebnisse der wissenschaftlichen Befragung (Schriftenreihe der Baden-Württemberg Stiftung Gesellschaft und Kultur Nr. 84). https://www.bwstiftung.de/de/publikation/nachhaltigkeit-lernen-ii-kinder-gestalten-zukunft. Zugegriffen: 28. Juni 2022

Impact Unit, Wissenschaft im Dialog (2021a) Entscheidungsbaum zur Evaluation von Wissenschaftskommunikation. https://impactunit.de/uebersicht-evaluationsplanung/. Zugegriffen: 28. Juni 2022

Impact Unit, Wissenschaft im Dialog (2021b) Evaluationsinstrumente entwickeln. (Wisskomm evaluieren). https://www.wissenschaftskommunikation.de/wp-content/uploads/2021b/09/How-To-4-Evaluationsinstrumente-entwickeln.pdf. Zugegriffen: 28. Juni 2022

Jackson M, Khalil K (2019) Expressions of empathy and related emotions towards animals: observational framework and code sheet (measuring empathy: a collaborative assessment project). Woodland Park Zoo, Oregon Zoo. https://resources.informalscience.org//expressions-empathy-and-related-emotions-towards-animals-observational-framework-and-code-sheet. Zugegriffen: 28. Juni 2022

Jahoda M, Lazarsfeld PF, Zeisel H (2021) Die Arbeitslosen von Marienthal: Ein soziographischer Versuch über die Wirkungen langandauernder Arbeitslosigkeit: mit einem Anhang zur Geschichte der Soziographie, 28. Aufl. Suhrkamp, Frankfurt a. M.

Knorr-Cetina K, Mulkay M (1983) Science observed. Perspectives on the social study of science. Sage, London

Kretschmer S (2017a) Analyse bisheriger empirischer Studien. In: Kretschmer S (Hrsg) Wissenschaft und Öffentlichkeit am Beispiel der Kinderuni: Theoretische Voraussetzungen und empirische Studien. Springer VS, Wiesbaden, S 161–195. https://doi.org/10.1007/978-3-658-15366-3_5

Kretschmer S (2017b) Empirische Studie Kinderuni Bonn. In: Kretschmer S (Hrsg) Wissenschaft und Öffentlichkeit am Beispiel der Kinderuni: Theoretische Voraussetzungen und empirische Studien. Springer VS, Wiesbaden, S 197–282. https://doi.org/10.1007/978-3-658-15366-3_6

Lamnek S, Krell C (2016) Qualitative Sozialforschung: Mit Online-Material. Beltz, Weinheim

Latour B, Woolgar S (1986) Laboratory life. The construction of scientific facts. Princeton University Press, Princeton

Lechner M (2018) Bildung für nachhaltige Entwicklung in botanischen Gärten: Entwicklung und Evaluation eines kompetenzorientierten Ausstellungskonzeptes zur Förderung von Perspektivenübernahme. Dissertation, Universität Tübingen. https://doi.org/10.15496/publikation-25848

Lublinski J (2004) Wissenschaftsjournalismus im Hörfunk: Redaktionsorganisation und Thematisierungsprozesse. UVK, Konstanz

Lüthje C (2017) Interne informelle Wissenschaftskommunikation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer VS, Wiesbaden, S 109–124. https://doi.org/10.1007/978-3-658-12898-2_6

Lux A, Theiler L (2021) Prozessbegleitende Evaluation von Kommunikations-und Partizipationsformaten im Themenfeld Bioökonomie. Evaluationskonzept BioKompass. ISOE-Materialien Soziale Ökologie, 66. Institut für sozial-ökologische Forschung, Frankfurt a. M.

Metten T, Bornheim F (2021) Responsive Wissenschaftskommunikation: Ein Realexperiment zur Bürgerbeteiligung in der Wissenschaftskommunikation, Katholische Universität Eichstätt-Ingolstadt. https://edoc.ku.de/id/eprint/25950/. Zugegriffen: 28. Juni 2022

Munsch M (2017) Konzeption und Evaluation eines Ausstellungsbereiches zum Thema „Evolutionäre Mechanismen". Dissertation, Universitäts- und Landesbibliothek Bonn https://bonndoc.ulb.uni-bonn.de/xmlui/handle/20.500.11811/7057. Zugegriffen: 28. Juni 2022

Nothhaft H (2011) Methode und Methodendiskussion—PR Research goes Management Research? In: Nothhaft H (Hrsg) Kommunikationsmanagement als professionelle Organisationspraxis: Theoretische Annäherung auf Grundlage einer teilnehmenden Beobachtungsstudie. VS Verlag für Sozialwissenschaften, Wiesbaden, S 115–191. https://doi.org/10.1007/978-3-531-92671-1_2

Pellegrini G (2021) Evaluating science communication: concepts and tools for realistic assessment. In Bucchi M, Trench B (Hrsg) Routledge handbook of public communication of science and technology, 3. Aufl. Routledge, New York

Pürer H (2009) Publizistik- und Kommunikationswissenschaft. UVK, Konstanz

Raupp J, Osterheider A (2019) Evaluation von Hochschulkommunikation. In: Fähnrich B, Metag J, Post S, Schäfer MS (Hrsg) Forschungsfeld Hochschulkommunikation. Springer VS, Wiesbaden, S 181–205. https://doi.org/10.1007/978-3-658-22409-7_9

Rennie L, McClafferty T (1996) Handbook for formative evaluation of interactive exhibits. Questacon – The National Science and Technology Centre, Canberra. https://resources.informalscience.org/sites/default/files/2013-11-22_Form_Eval_handbook_Rennie%26McClaff_1996.pdf. Zugegriffen: 28. Juni 2022

Schäfer MS, Kristiansen S, Bonfadelli H (2015) Wissenschaftskommunikation im Wandel: Relevanz, Entwicklung und Herausforderungen des Forschungsfeldes. In: Schäfer MS, Kristiansen S, Bonfadelli H (Hrsg) Wissenschaftskommunikation im Wandel. Herbert von Halem, Köln, S 10–42

Schmitt-Sody B (2013) NESSI-FÖSL. Konzeption und Evaluation eines Schülerlabors für Förderschüler aus chemiedidaktischer Perspektive. Dissertation, Friedrich-Alexander-Universität, Erlangen-Nürnberg

Thierbach C, Petschick G (2019) Beobachtung. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer VS, Wiesbaden, S 1165–1181. https://doi.org/10.1007/978-3-658-21308-4_84

Trauner S (2016) „Komm, mach mit!" Oder besser: „Bitte stör' mich nicht!"? In: Decristofero B, Hopmann S, Kartschnig T, Seebauer L, Swertz C, Decristoforo B, Katschnig T (Hrsg) Hands-On im Technischen Museum Wien: Konzeption und Evaluierung der Mitmachausstellung „In Bewegung", 1. Aufl. nap, new academic press, Wien, S 71–81

Wehmeier S, Raaz O (2016) Nicht standardisierte Methoden in der PR-Forschung. In: Averbeck-Lietz S, Meyen M (Hrsg) Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft. Springer VS, Wiesbaden, S 415–427. https://doi.org/10.1007/978-3-658-01656-2_30

Ziegler R, Hedder IR, Fischer L (2021) Evaluation of science communication: current practices, challenges, and future implications. Front Commun 6. https://doi.org/10.3389/fcomm.2021.669744

André Weiß ist wissenschaftlicher Mitarbeiter am Department für Wissenschaftskommunikation des Karlsruher Institut für Technologie (KIT). Neben Lehrtätigkeiten in den Studiengängen Wissenschaft – Medien – Kommunikation (B.A. und M.A.) erarbeitet er als Dissertation eine qualitative Analyse von biographischen Wendepunkten und deren Bedeutung für die Rezeption von Wissenschaftskommunikation.



Grundlagenbeitrag: Nutzungsdatenanalyse digitaler Medien als Instrument der evaluativen Wissenschaftskommunikationsforschung

Armin Hempel

Zusammenfassung

Der Grundlagenbeitrag skizziert die Nutzungsdatenanalyse für den Bereich der Evaluationsforschung digitaler Wissenschaftskommunikationsprojekte und diskutiert ihre Stärken wie ihre Herausforderungen. Dazu werden verschiedene Typen von Nutzungsdaten beschrieben und exemplarisch Werkzeuge vorgestellt, diese zu erheben. Inwiefern kann es lohnenswert sein, Nutzungsdatenanalyse innerhalb von digitalen Wissenschaftskommunikationsprojekten zu betreiben und was kann bereits während der Projektplanungsphase beachtet werden, um Nutzungsdatenanalysemethoden sinnvoll einzusetzen? Es wird beschrieben, wie Erwägungen zum Datenschutz eine Herausforderung für die Nutzungsdatenanalyse darstellen, der damit verbundene Aufwand jedoch durch die Verwendung passender Werkzeuge auf ein Minimum reduziert werden kann. Zuletzt wird das Desiderat formuliert, die Nutzungsdatenanalyse als Methodenset weiter zu elaborieren – auch, um zukünftige digitale Wissenschaftskommunikationsformate untersuchen zu können.

A. Hempel (✉)
SFB 980 „Episteme in Bewegung“, Freie Universität Berlin, Berlin, Deutschland
E-Mail: armin.hempel@fu-berlin.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_10

1 Einleitung

Die Digitalisierung transformiert unsere Gesellschaften seit Mitte der 1990er Jahre grundlegend. Diese Transformation erstreckt sich über alle gesellschaftlichen Teilbereiche und verändert so auch die Wissenschaften sowie deren Beziehung zur Gesellschaft nachhaltig (Neuberger et al. 2021).

Gleichzeitig erfahren auch die Wissenschaftskommunikation und deren Erforschung eine Transformation: Jedes neu entstehende digitale Medium bietet immer auch eine neue Möglichkeit, Wissenschaft zu kommunizieren. Damit einher gehen dann immer auch neue Forschungsfragen, die es erforderlich machen, diese digitalen Wissenschaftskommunikationsformen zu untersuchen. Das umfasst Fragen nach den untersuchten Projekten an sich, den Zielen, die mit ihnen verfolgt werden, ihrer Wirkung sowie danach, wie sie zu verbessern sind (Ziegler und Fischer 2020).

Beim Nachdenken über evaluative Forschungsdesigns im Rahmen der Wissenschaftskommunikationsforschung liegt es aus unterschiedlichen Gründen nahe, Methoden der Nutzungsdatenanalyse mit einzubeziehen.

Denn – egal, ob Social-Media-Präsenzen, Blogs, Podcasts, Smartphone-Apps, Internetvideos, E-Mail-Newsletter, Messenger-Dienste oder klassische Projektwebseiten – bei so gut wie allen digitalen Medien der Wissenschaftskommunikation werden automatisch Nutzungsdaten generiert, die für die Bewertung von Projekten als unterstützende Datengrundlage herangezogen werden können.

Nutzungsdaten können dabei – je nach Medium und Auslegung – bloße Abrufzahlen, aber auch personenbezogene Daten sein. Sie erlauben Rückschlüsse auf den geografischen Aufenthaltsort und die technische Ausstattung der Nutzer:innen oder enthalten Informationen zur Verweildauer und dem eingeschlagenen Weg durch eine Website. Häufig bleiben diese Daten ungenutzt, sei es aus Unkenntnis darüber, dass sie existieren, aufgrund von Schwierigkeiten, auf sie zuzugreifen bzw. sie zu interpretieren oder auch aus Sorge vor datenschutzrechtlichen Problemen.

Dieser Beitrag richtet sich an Forschende und Praktiker:innen – explizit an Einsteiger:innen in das Gebiet – die planen, im Rahmen eines evaluativen (Forschungs)-Designs ihr eigenes Projekt unter die Lupe zu nehmen und selbst Nutzungsdatenanalyse zu betreiben. Er versteht sich als eine erste Handreichung zur Annäherung an das Themenfeld. Die Nutzungsdatenanalyse wird dabei nicht als eine etablierte wissenschaftliche Methode vorgestellt, sondern es werden verschiedene Perspektiven auf die Nutzungsdatenanalyse als Instrument skizziert und ein Überblick über ihre Stärken und Schwächen gegeben.

Der detailliertere Umgang mit den durch Nutzungsdatenanalyse erlangten Daten, ihre Bereinigung, Verarbeitung, Aufbereitung, Visualisierung, Archivierung und eventuelle Weitergabe können nicht Gegenstand dieses Beitrags sein und müssen unberücksichtigt bleiben. Es existiert aber eine Vielzahl von disziplinspezifischen Leitlinien zum Forschungsdatenmanagement (Deutsche Forschungsgemeinschaft [DFG] 2021), die hier weiterhelfen können.

2 Nutzungsdatenanalyse – Eine Kurzdefinition

Nutzungsdatenanalyse ist die Sammlung, systematische Erhebung und Auswertung von selbst oder durch Dritte erhobenen Daten, die bei der Nutzung digitaler Angebote anfallen (McFadden 2005). Ihre Verfahren werden auch mit den Begriffen „Logfile-Analyse“, „Web Mining“, „User Tracking“ oder „Social-Media-Analyse“ umschrieben (Priemer 2004). Während sie in der Marktforschung als etabliertes Werkzeug gilt, um bspw. zu erfahren, ob eine Werbekampagne erfolgreich ist, wie gut eine Website funktioniert, welche Zielgruppen intensiver umworben oder fallen gelassen werden sollten oder in welcher Art ein eCommerce-Angebot gestaltet sein muss, um letztendlich viele Verkäufe zu realisieren, ist der Einsatz der Nutzungsdatenanalyse in anderen Feldern nicht unumstritten. Im Online-Journalismus beispielsweise wird die unreflektierte Anwendung von Nutzungsdatenanalysen dafür kritisiert, dass redaktionelle Entscheidungen darauf fußen, wie und mit welchen Inhalten die meisten Klicks generiert werden können, um Werbeumsätze in die Höhe zu treiben. So könne z. B. eines der Ziele des Journalismus, das Befördern des demokratischen Prozesses, durch eine rein auf Klickzahlen ausgelegte Publikationsstrategie in Gefahr gebracht werden (Tandoc und Thomas 2015).

Ähnlich gelagerte, grundlegende Divergenzen zwischen den beiden Forschungskulturen der akademischen Sozialforschung und den Methoden der Marktforschung beschreibt Ziegler (2014) detailliert.

Aus der Betrachtung dieser Vorbehalte ergibt sich, dass die Techniken der Nutzungsdatenanalyse nur insoweit für die evaluative Wissenschaftskommunikationsforschung Anwendung finden sollten, als sie ethische Grundsätze, insbesondere die des Datenschutzes, die Standards der Evaluationspraxis im Hinblick auf Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit (siehe auch van den Bogaert in diesem Band), und – ganz allgemein – die Grundsätze der guten wissenschaftlichen Praxis nicht verletzen (DFG 2019).

Wenn in diesem Artikel von Nutzungsdatenanalyse die Rede ist, ist damit ausdrücklich nicht das *Data Scraping* gemeint, wie es z. B. von Batrinca und

Trelaeven (2015) beschrieben wird: Ein Verfahren, das Webseiten von Dritten oder ganze soziale Netzwerke in den Blick nimmt und das häufig eine Rolle für die Inhaltsanalyse dieser Seiten spielt. Nutzungsdatenanalyse im Sinne dieses Artikels schließt auch andere, sogenannte Offsite-Tools aus, die ein Wissenschaftskommunikationsprojekt aus einer Außenperspektive heraus untersuchen.

Außerdem soll es nicht um das sogenannte *Web Tracking* gehen, das versucht, mit Techniken wie dem *Browser Fingerprinting* (Boda et al. 2012) ein möglichst detailliertes Bild einer einzelnen Person zu zeichnen, indem ihre Aktivitäten beim Besuch vieler verschiedener Webseiten und über einen möglichst langen Zeitraum hinweg aufgezeichnet und ausgewertet werden, um z. B. ihr Konsumverhalten möglichst präzise vorhersagen zu können. Insbesondere aufgrund solcher – datenschutzrechtlich eher zweifelhafter – Praktiken steht die Nutzungsdatenanalyse häufig in der Kritik.

Der vorliegende Beitrag begrenzt sich auf Instrumente der sogenannten *Reichweitenanalyse* (genauer beschrieben durch Fritz 2004), die versuchen, aus der Perspektive eines einzelnen digitalen Angebots heraus Informationen über dessen Interaktionen mit Besucher:innen und das technische Funktionieren dieses Angebots zu erlangen.

Eine Reichweitenanalyse kann Aufschluss darüber geben, wie viele Besucher:innen ein Angebot pro Tag hat, darüber, was die populärsten Episoden, Unterseiten oder Beiträge eines Projekts der digitalen Wissenschaftskommunikation sind, wie viel Zeit Besucher:innen mit dem Angebot verbringen, wie sie zum Projekt gefunden haben, welche Sprache sie sprechen und vieles mehr (Clifton 2012). Die Reichweitenanalyse bietet die Basis für evaluative Fragestellungen darüber, welche Teile eines Angebots gut funktionieren oder welche das meiste Verbesserungspotenzial bieten, und spielt ihre Stärken insbesondere dann aus, wenn sie mit anderen Methoden kombiniert wird, wie bereits durch Welker und Wünsch (2010) vorgeschlagen. So wären beispielsweise Fragestellungen zum Zusammenhang der sprachlichen Komplexität einzelner Textbeiträge mit ihrer Nutzungshäufigkeit in spezifischen Zielgruppen denkbar.

Die folgenden Unterkapitel beschreiben, wo und wie Nutzungsdaten anfallen, was für und was gegen eine Anwendung von Werkzeugen der Nutzungsdatenanalyse in der evaluativen Wissenschaftskommunikationsforschung spricht und was es bei der Planung eines digitalen Wissenschaftskommunikationsprojekts zu beachten gilt, wenn die Nutzungsdatenanalyse eine Rolle bei seiner evaluativen Betrachtung spielen soll.

3 Was sind Nutzungsdaten?

Nutzungsdaten fallen durch die Nutzung eines digitalen Angebots, z. B. beim Besuch einer Internetseite, an. Dies können zum einen *Zielgruppendaten* sein, also solche, die Eigenschaften der Nutzer:innen einer digitalen Wissenschaftsk ommunikationsmaßnahme beschreiben, zweitens *Aktivitätsdaten,* also solche, die die Aktivitäten der Nutzer:innen während ihres Besuchs betreffen oder drittens *technische Daten,* die Aufschluss über technische Aspekte eines digitalen Angebots geben (Guba und Gebert 1998).

Beispiele für *Zielgruppendaten* im Rahmen einer Reichweitenanalyse eines Internetangebots sind Angaben über die geografische Region, aus der ein Zugriff erfolgt, die im Webbrowser eingestellte Sprache und darüber, mit welchem Endgerät, unter Verwendung welchen Betriebssystems, welchen Browsers und in welcher Bildschirmauflösung auf ein Angebot zugegriffen wurde.

Aktivitätsdaten beinhalten die Dauer eines Besuchs, nach welchen Begriffen auf einer Seite gesucht wurde, welche Wege die Besucher:innen durch die Webseite genommen haben (Funnel Analytics), über welchen Weg und zu welcher Uhrzeit Besucher:innen zum Angebot fanden oder an welcher Stelle ein Angebot wieder verlassen wurde.

Technische Daten umfassen die Anzahl der Besuche einer Internetseite innerhalb eines bestimmten Zeitraums und die Information darüber, welche Unterseiten wie häufig besucht wurden, die Zeit, die eine Webseite im Schnitt benötigt, um sich aufzubauen, die Häufigkeit und Typen bestimmter Server-Fehlermeldungen (z. B. „404 – Seite nicht gefunden“). Außerdem zählen Daten, die über die Auslastung des Servers in Bezug auf dessen benötigte Rechenleistung oder Serverkapazität Aufschluss gewähren, zu den technischen Daten.

Nutzungsdaten fallen entweder über die Anwendung sogenannter *Page Tags* – meist gestützt durch die Verwendung von Cookies – im Browser der Nutzer:innen oder als Log-Dateien direkt auf einem Webserver an. Ein Eintrag in solch einem *Server Log,* genauer spezifiziert in einem Working Draft des World Wide Web Consortiums (Hallam-Baker und Behlendorf 1996), kann folgendermaßen aussehen:

```
94.130.145.107 - - [20/Jan/2022:15:37:21 +0100] "GET
/transkripte/Beispiel.pdf HTTP/1.1" 206 16384
"https://www.wissenschaftskommunikation.de/"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)
AppleWebKit/605.1.15 (KHTML, like Gecko) Version/15.2
Safari/605.1.15"
"wissenschaftskommunikationsevaluationsforschung.de"
```

Der Eintrag besteht aus den Komponenten *IP-Adresse des Clients* (des/der Besucher:in) [94.130.145.107]; *Identität,* findet in der Praxis meist keine Anwendung [-]; *Benutzername*, falls die abgerufene Datei durch ein Passwort geschützt ist [-]; *Zeitstempel* [20/Jan/2022:15:37:21+0100]; *Anfrage des Clients,* untergliedert in Methode, Dateiname und verwendetes Protokoll [GET/transkripte/Beispiel.pdf HTTP/1.1]; *HTTP-Statuscode* der Server-Antwort [206]; *Dateigröße* der ausgelieferten Datei (in Bytes) [16384]; *HTTP-Referrer* (die Website, über die ein:e Besucher:in zur aktuellen Seite gekommen ist); *User Agent,* also Informationen über den für den Abruf verwendeten Browser und das Betriebssystem [Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/15.2 Safari/605.1.15]; *Abgerufenes Objekt* [wissenschaftskommunikationsevaluationsforschung.de].

Log-Einträge in dieser oder ähnlicher Form legt jede Webserver-Software automatisch an. Die Einträge können entweder direkt verarbeitet oder durch eine Analyse-Software mit Page Tags oder anderen Daten kombiniert, aufbereitet und miteinander verknüpft werden[1] und bieten so einen schnellen Einblick in komplexere Zusammenhänge – denkbar ist z. B. ein Vergleich der Anzahl der Abrufe mehrerer Podcast-Episoden in den ersten zwei Monaten nach ihrer jeweiligen Veröffentlichung. Je nach genutztem Medium können aber auch vollkommen andere Datentypen und Visualisierungsarten zur Verfügung stehen, z. B. so genannte Heatmaps, die Aufschluss über häufig geklickte Bereiche auf einer Website geben oder Daten darüber, welche Web-Videos zu besonders vielen Abonnements des eigenen Kanals geführt haben. Eine genauere Analyse der Vor- und Nachteile des Zusammenwirkens von Page Tags, Cookies, Server Logs sowie anderer, weniger verbreiteter Verfahren liefert Clifton (2012).

Nachfolgend wird ein Überblick über die wichtigsten Typen von Nutzungsdaten (auch: Metriken) gegeben, die in den unterschiedlichen Medien der digitalen Wissenschaftskommunikation anfallen können (Tab. 1):

Abhängig von der Fragestellung oder dem Forschungsinteresse kann es entweder sinnvoll sein, einige dieser Werte miteinander zu verknüpfen oder sich auf nur auf einzelne Werte zu konzentrieren. Üblicherweise werden in den gängigen Analyseinstrumenten Besuchszeiten, Wege durch die Webseite oder Akquisitionsdaten, die Auskunft über die Website geben, von der ein:e Besucher:in kommt, visualisiert.

[1] Üblich ist z. B. eine Standortdatenbank, die Geolocations zu IP-Adressen bereithält.

Tab. 1 Typen von Nutzungsdaten mit Relevanz für digitale Wissenschaftskommunikationsprojekte

Wissenschaftskommunikationsprojekte auf Web-Basis (Webseiten, Web-Apps, Blogs, Podcasts)
Anzahl der Seitenbesuche, Downloads bzw. Hörintentionen, aufgerufene URL, Seitentitel, Aufenthaltsdauer bzw. geschätzte Wiedergabezeit, Zugriffsquellen (der Weg, über den Besucher:innen zum Inhalt gelangen, ob über andere Websites, durch Suchen, durch Social Media etc.), Geolocations, Datum und Uhrzeit des Aufrufs, Gerätetyp, -modell, -hersteller, Bildschirmauflösung und verwendetes Betriebssystem, Absprungrate, Aktionen pro Besuch, Anzahl und Inhalte der Suchen, Anzahl, Inhalte und Dateigrößen von Downloads, Ort der Wiedergabe (Download, RSS-Feed oder Podcast-Web-Player) Daraus abgeleitet: Der Weg durch die Webseite; Bedingt (im Falle von externem Hosting): Andere Datentypen, z. B. Blogs oder Podcasts, die die Benutzer:innen sehen, lesen oder hören
E-Mail-Newsletter:
Anzahl Empfänger:innen, Öffnungsrate, Datum und Uhrzeit der Öffnung der E-Mail, Klicks auf Links, Rückläufer, Abmeldungen, Geolocations, bedingt: Gerätetyp
Webvideos:
Anzahl der Aufrufe, Gesamtwiedergabezeit, Zugriffsquellen, Geolocations, Datum und Uhrzeit des Aufrufs, Gerätetyp und verwendetes Betriebssystem, Ort der Wiedergabe (direkt oder eingebettet), bedingt: Untertiteltyp, bedingt (nur bei externem Hosting): Alter und Geschlecht der Zuschauer:innen, andere Kanäle und Videos, die sich die Zuschauer:innen angesehen haben
Twitter (bezogen jeweils auf ein einzelnes Posting):
Veröffentlichungszeitpunkt, Impressions (Sichtkontakte), Interaktionen, Interaktionsrate, Retweets, Antworten, „Gefällt mir"-Angaben, Nutzerprofilklicks, URL-Klicks, Hashtag-Klicks, Detailerweiterungen, Permalink-Klicks, Follows, Medienanzeigen, Medieninteraktionen
Facebook (bezogen jeweils auf ein einzelnes Posting):
Veröffentlichungszeitpunkt, Reichweite, „Gefällt mir"-Angaben und Reaktionen, Antworten, Link-Klicks, Kommentare, Geteilte Inhalte, grobe Altersverteilung und Wohnorte der Zielgruppe
Instagram (bezogen jeweils auf ein einzelnes Posting):
Profilaufrufe, Reichweite, Interaktionen, neue Abonnent:innen, Impressions, „Gefällt mir"-Angaben, Anzahl Kommentare, Geteilte Inhalte, Favorisierte Inhalte

4 Warum Nutzungsdatenanalyse (in der evaluativen Wissenschaftskommunikationsforschung) betreiben?

Bereits in der Planungsphase eines Wissenschaftskommunikationsprojekts stellt sich häufig die Frage nach den Zielen und den Zielgruppen eines Projekts, und es kommen Fragen danach auf, welche Evaluationsmethode sich eignet, um das Erreichen dieser Ziele zu überprüfen. Eine Übersicht über mögliche Motive und Ziele in der Wissenschaftskommunikation und konkret in Hinblick auf deren Evaluation geben Ziegler und Fischer (2020).

Bei *digitalen* Wissenschaftskommunikationsprojekten kann bereits ein regelmäßiger Blick auf die Nutzungsdatenauswertung dabei helfen, das eigene Projekt zu reflektieren sowie ein Gespür für seine Reichweite, insbesondere für seinen Output zu bekommen. Eine Frage, die sich an jedes Wissenschaftskommunikationsprojekt stellen lässt, ist die danach, ob das Projekt für die eigene Zielgruppe optimal erreichbar ist. Hier kann z. B. ein Web-Analysetool wie *Matomo* (ehemals *Piwik*) dabei helfen, Schwachstellen und Fehler in der Webseite des Projekts zu identifizieren, oder einzelne Unterseiten ausfindig zu machen, die langsam oder überhaupt nicht geladen werden können. Zugangsbarrieren bei der Erreichbarkeit der Seite müssen dabei nicht zwangsläufig technischer Natur sein, auch ein zu komplexes oder verwirrendes User-Interface kann durch Nutzungsdatenanalyse als Fehlerquelle entlarvt werden. Eine kurzweilige und leicht verständliche Einführung ins User-Interface-Design bietet Steve Krug in „Don't make me think!“ (2014). Je nachdem, welche Beweggründe einem Evaluationsvorhaben zugrunde liegen, liefert die Nutzungsdatenanalyse einen Beitrag zum Erkenntnisgewinn. Mögliche Beweggründe sind der Wunsch danach, mehr Wissen über das Wissenschaftskommunikationsprojekt zu erlangen, Optimierungspotenziale aufzudecken oder Ergebnisse und Wirkungen des Projekts zu prüfen (siehe Impact Unit 2021). Je nach Evaluationsdesign kann die Nutzungsdatenanalyse sowohl als Komponente einer formativen Evaluation oder einer summativen Evaluation dienen (siehe auch Volk in diesem Band). Insbesondere für Evaluationsdesigns, deren Ziel es ist, ein Projekt noch während seiner Laufzeit im Prozess zu verbessern, bietet sich die Nutzungsdatenanalyse an.

Stellen wir uns folgendes Beispiel vor: Die anvisierte Zielgruppe einer naturwissenschaftlich ausgerichteten YouTube-Video-Reihe ist Schüler:innen zwischen 13 und 18 Jahren. Schon jeweils kurz nach Veröffentlichung der ersten drei

Videos stellt sich beim Blick in die Statistiken heraus, dass fast ausschließlich Zuschauer:innen über 65 Jahre erreicht werden. Dadurch bietet sich die Gelegenheit, im Produktionsprozess weiterer Videos frühzeitig gegenzusteuern – oder das Projekt und seine Zielsetzung zu überdenken.

Viele Logfile-Analysen haben zum Ziel, Typen von Nutzer:innen und verschiedene Strategien, mit denen diese Nutzer:innen versuchen, bestimmte Informationen zu erlangen, zu identifizieren und sie mit weiteren Variablen oder bestimmten Konstrukten wie Lernerfolgen, die in Befragungen erhoben werden, in Zusammenhang zu bringen (Priemer 2004). Innerhalb der Wirkungsforschung sind beispielsweise Fragestellungen dazu vorstellbar, inwiefern sich das Bild von Wissenschaft im Allgemeinen oder von einer Disziplin im Speziellen für Nutzer:innen im Laufe der Zeit ändert, wenn diese regelmäßig bestimmte Podcasts hören oder einen Blog lesen. Denkbar wäre zur Annäherung an diese Fragestellung eine Kombination von Nutzungsdatenanalyse mit qualitativen und quantitativen Befragungsinstrumenten. Auch Fragen nach dem spezifischen Interesse von Besucher:innen als Grund für den Besuch eines Angebots lassen sich unter Zuhilfenahme der Nutzungsdatenanalyse stellen – indem man z. B. verwendete Suchbegriffe betrachtet.

Fagan (2013) beschreibt, wie die Adaption von bereits etablierten Leistungskennzahlen *(Key Performance Indicators, KPIs)* aus dem Marketing für das akademische Bibliothekswesen fruchtbar gemacht werden kann. Besonders im Hinblick auf jene *KPIs,* die das *User Engagement* betreffen, könnte eine solche Adaption auch auf Fragestellungen der Wissenschaftskommunikationsforschung anwendbar sein, z. B. auf solche, anhand derer die Einbindung und die Beteiligung von Bürger:innen an Forschungsprozessen untersucht werden sollen.

Nutzungsdatenanalyse kann für die Evaluation von Wissenschaftskommunikation aber nicht nur eine Methode zur Generierung von Forschungsdaten sein und somit zur Verbesserung einer Wissenschaftskommunikationsm aßnahme beitragen. Bereits das bloße Nachdenken über die Möglichkeiten der Nutzungsdatenanalyse kann bei der Generierung neuer Forschungsfragen unterstützen.

Im Praxisbeitrag zur Nutzungsdatenanalyse in diesem Band (siehe auch Buckermann und Greving in diesem Band) wird die Untersuchung von Beteiligungsmustern an einem Bürgerwissenschaftsprojekt beschrieben, mit dem Ziel, Aussagen über die Motivation der Teilnehmenden treffen zu können. Voraussetzung für das Stellen einer solche Forschungsfrage ist die Kenntnis über die Möglichkeiten der Reichweitenanalyse, speziell der Auswertung der Aktivitätsdaten von Webseiten-Besucher:innen.

Eines der gewichtigsten Argumente für die Erwägung von Nutzungsdatenanalyse im Rahmen der Evaluation von Wissenschaftskommunikationsmaßna hmen ist aber der geringe Aufwand, der mit ihrer Durchführung verbunden ist. Bei beinahe jedem Vorhaben der digitalen Wissenschaftskommunikation stehen Nutzungsdaten mehr oder weniger automatisch zur Verfügung – ohne aufseiten der Forscher:innen oder der Nutzer:innen einen zusätzlichen Zeitaufwand zu erzeugen. Die Erhebung von Nutzungsdaten ist meist leicht einzurichten, wenig aufwendig während der Durchführung und außerdem kostengünstig zu betreiben. Nutzungsdatenanalyse erfordert kein Experimentaldesign, keinen Entwurf von Fragebögen und kann – einmal eingerichtet – automatisiert objektive, eindeutige, vollständige und detailgenaue Daten (Priemer 2004) über die komplette Projektlaufzeit hinweg liefern.

Darüber hinaus ist sie nicht-invasiv, sie erfordert vonseiten der Nutzer:innen eines Angebots der digitalen Wissenschaftskommunikation keinerlei Ressourcen, weder deren Zeit noch ihre besondere Aufmerksamkeit – Nutzungsdaten werden „nebenbei" erhoben, während die Besucher:innen eines Angebots dieses nutzen – ohne, dass sie gestört werden.

Eine der Herausforderungen bei der Anwendung der Nutzungsdatenanalyse als Instrument der evaluativen Wissenschaftskommunikationsforschung ist, dass es in manchen Fällen sehr schwierig sein kann, die erforderlichen Daten zu erlangen. Dies gilt insbesondere für Daten, die aus den proprietären Analysewerkzeugen sozialer Netzwerke gewonnen werden müssen. Außerdem wird ein nicht unerheblicher Teil von Server-Log-Dateien durch automatisieren Traffic (Suchmaschinen-Bots etc.) erzeugt und muss häufig herausgefiltert werden, da mittels dieser Daten keine Erkenntnisse über das Verhalten von Nutzer:innen gewonnen werden können. Ohne eine Kombination mit anderen Methoden, wie z. B. mit Befragungen oder Inhaltsanalysen, sind die Ergebnisse der Nutzungsdatenanalyse nur bedingt aussagekräftig (Priemer 2004). Nicht zuletzt können auch datenschutzrechtliche Erwägungen, insbesondere jene mit Blick auf die 2016 eingeführte europäische Datenschutzgrundverordnung (DSGVO)[2], die Möglichkeiten und die Aussagekraft von Nutzungsdatenanalyse einschränken.

[2] Eine ausführlichere Aufschlüsselung der Herausforderungen datenschutzrechtlicher Erwägungen im konkreten Projektkontext in Abschn. 5.3.

5 Was ist bei der Gestaltung eines Wissenschaftskommunikationsprojekts zu beachten?

Grundsätzlich bieten alle digitalen Wissenschaftskommunikationsprojekte die Möglichkeit, Nutzungsdatenanalyse zu betreiben. Das gewählte Medium und die konkrete Ausgestaltung eines Projekts bestimmen dabei die Herangehensweise, die Daten zu erheben und beeinflussen damit auch die Forschungsfragen, die gestellt werden können.

5.1 Einfluss des gewählten Mediums auf Möglichkeiten der Nutzungsdatenanalyse

Das gewählte Medium hat einen großen Einfluss darauf, welche Typen von Daten auf welche Art und Weise erhoben werden können. Während für Websites, Web-Apps, Blogs, E-Mail-Newsletter-Systeme und Podcasts, also für Medien, die grundsätzlich durch einen eigenen Webserver bereitgestellt werden können, eine breite Palette von Analysewerkzeugen und -techniken verfügbar ist, ist der Zugriff auf die Nutzerdaten plattformbasierter oder sozialer Medien wie Twitter, Instagram, Facebook, Whatsapp oder YouTube sowie bei nativen Smartphone-Apps immer auf die Analysewerkzeuge begrenzt, die die Anbieter:innen dieser Plattformen und App-Stores zur Verfügung stellen.

Die qualitativen Unterschiede zwischen diesen Werkzeugen sind hoch. Twitter Analytics beispielsweise ist ein leicht zugängliches und klar strukturiertes Analysewerkzeug, das immerhin einen direkten Datendownload im csv-Format erlaubt. Ähnlich – wenn auch eingeschränkter – verhält es sich bei Facebook Insights. Bei Instagram hingegen ist der Insights-Bereich nur aus der mobilen App heraus und nur für spezielle Business- oder Creator-Accounts zugänglich, auch ein direkter Datendownload ist nicht möglich. Die Nutzungsdaten manch anderer sozialer Netzwerke sind sogar gar nicht oder nur dann zugänglich, wenn eine spezielle Drittanbieter-Software für die Datenauswertung genutzt wird. Nichtsdestotrotz können sich auch Daten aus schwer zugänglichen Quellen für eine Nutzungsdatenanalyse in der evaluativen Wissenschaftskommunikationsforschung eignen: Beispielsweise stützen Essig et al. (2020) ihre Analyse eines Instagram-Accounts als geeignetes Mittel, um histologisch Befunde für Medizinstudent:innen zugänglich zu machen einerseits auf Befragungen, andererseits aber auch auf erhobene Nutzungsdaten.

5.2 Einfluss des Hosting-Providers auf die Möglichkeiten der Nutzungsdatenanalyse

Während der Planung eines digitalen Wissenschaftskommunikationsprojektes stellt sich häufig die Frage danach, wie und durch wen die produzierten Inhalte zur Verfügung gestellt werden sollen; ob es beispielsweise sinnvoller ist, sie über einen eigenen Webserver oder durch einen externen Blog-, Webvideo- oder Podcast-Hosting-Anbieter bereitzustellen. Bei der Entscheidung darüber können technische, datenschutzrechtliche oder Erwägungen zur Barrierefreiheit und Zugänglichkeit eine Rolle spielen. Ebenso stellen sich Fragen danach, wie aufwendig ein eigenes Hosting einzurichten und zu pflegen ist, oder welche Kosten bei der Nutzung einer externen Plattform entstehen.

Auch Fragen nach der Verfügbarkeit, der jeweiligen Aufbereitung und den Downloadmöglichkeiten von Nutzungsdaten müssen abgewogen werden. Wird ein eigener Server für das Hosting des Wissenschaftskommunikationsprojekts genutzt, bringt das einerseits den Vorteil mit sich, dass der unbeschränkte und wiederholte Zugriff auf die Server-Logdateien möglich ist. Darüber hinaus besteht die Möglichkeit, unter vielen verschiedenen Datenanalyse-Tools zu wählen. Auch diese Tools lassen sich dann entweder auf dem eigenen Webserver installieren – beispielsweise *AWStats* oder *Webalizer* – oder als Software-as-a-Service bei einem Drittanbieter einkaufen, wie es z. B. für den Einsatz von *Google Analytics* zutrifft. Eine selbstgehostete Analyse-Software läuft meist auf demselben Webserver wie die zu analysierende Website, sei es als Plugin im verwendeten Content-Management-System oder als separate, eigenständige Software. Die freie Analyse-Software *Matomo* beispielsweise bietet mehr als 100 Plugins, die zur Vereinfachung ihrer Einbindung in verschiedenste Content-Management-Systeme dienen. Ansätze systematischer Reviews der verbreitetsten Webanalyse-Tools finden sich in Bekavac und Garbin Praničević (2015) sowie in Kumar und Ogunmola (2020).

Wird dagegen ein digitales Wissenschaftskommunikationsprojekt von einem externen Anbieter gehostet, geht dadurch die Entscheidungshoheit darüber verloren, welche Nutzungsdaten erhoben werden und in welcher Art diese Daten dargestellt werden. Die proprietären Analyse-Werkzeuge der Hosting-Anbieter sind häufig in ihrem Datenzugriff, in ihren Analysefähigkeiten sowie in der Weise, in der die Daten visualisiert werden, limitiert. Sie können aber – bei allen Einschränkungen – auch Einblicke bieten, die mit einer selbstgehosteten Analyse-Software nicht zu erlangen sind. Einige Podcast-Hosts stellen z. B. Informationen darüber zur Verfügung, welche anderen Podcasts bzw. Podcast-Kategorien

Hörer:innen des eigenen Podcasts hören oder liefern eine minutengenaue Übersicht über die Anzahl der Hörer:innen pro Episode. Der Video-Host *YouTube* bietet eine Übersicht über Alter und Geschlecht der Zuschauer:innen oder die Uhrzeit, zu der sie üblicherweise Videos schauen. Das sind Informationen, die sich weder aus einer Auswertung der Server-Log-Dateien noch über JavaScript-Tracking ergeben können. Diese Daten sind nur dann zu erlangen, wenn ein Hosting-Dienst auf ergänzende Daten zurückgreift, über die ausschließlich dieser selbst verfügt. Dies können persönliche Daten wie z. B. Alter und Geschlecht oder Nutzungsdaten der Endgeräte sein, mit denen Podcasts gehört oder Videos gesehen werden. Im Podcast-Bereich erscheint das sogenannte *Client-Side-Tracking,* also das Einbeziehen von Nutzungsdaten, die im Podcatcher (also direkt auf dem jeweiligen Abspielgerät), erhoben werden, präzisere Daten zu versprechen, es ist allerdings noch nicht sehr verbreitet (Podigee 2019).

Zusammenfassend lässt sich also sagen, dass es sich aus der Perspektive der evaluativen Wissenschaftskommunikationsforschung eher empfiehlt, ein digitales Wissenschaftskommunikationsprojekt unter Verzicht auf externe Anbieter:innen zu hosten, um in der Wahl der Analysewerkzeuge flexibel zu bleiben und einen direkten Zugriff auf Server-Log-Dateien zu haben. Nur in Ausnahmefällen sind Nutzungsdaten, die ausschließlich durch externe Hosting-Dienste angeboten werden, so interessant, dass ein Fremdhosting dem eigenen Server vorzuziehen ist. Sollte trotzdem ein solches Fremdhosting in Betracht gezogen werden, kann es hilfreich sein, den Hosting-Anbieter um die regelmäßige Bereitstellung der Server-Log-Dateien für das eigene Angebot zu bitten.

5.3 Erwägungen zum Datenschutz

Nutzungsdaten sind häufig personenbezogene Daten, die zur Identifikation oder Standortbestimmung einzelner natürlicher Personen genutzt werden können. Auch, wenn die Nutzungsdatenanalyse zum Zwecke der wissenschaftlichen Evaluation oder der Verbesserung einer Wissenschaftskommunikationsmaßnahme durchaus häufig als „wissenschaftlicher Forschungszweck" nach Artikel 9 Abs. 2 der DSGVO beurteilt werden kann, ist es erforderlich, nur so viele personenbezogene Daten wie unbedingt nötig zu erheben. Weiterhin muss sichergestellt sein, dass nicht mehr Personen als unbedingt nötig Zugriff auf die erhobenen Daten erhalten und dass diese Daten nicht länger als nötig gespeichert werden. Darüber hinaus ist es ratsam, in der eigenen Datenschutzerklärung so transparent wie möglich darüber zu sein, welche Daten zu welchen Zwecken gesammelt werden und wie lange diese Daten gespeichert werden. Auch die Frage danach,

ob eine Datenschutz-Folgenabschätzung oder ein Datenschutzkonzept erstellt werden muss, sollte in der Planungsphase eines Projekts immer mit dem/der zuständigen Datenschutzbeauftragten besprochen werden.

Proprietäre (nichtsdestotrotz populäre) Webanalyse-Tools von Anbietern wie Google oder Adobe werden in Bezug auf den Datenschutz häufig kritisch gesehen – unter anderem, weil bei ihrer Nutzung nicht zweifelsfrei geklärt werden kann, auf welchen Servern die Daten verarbeitet werden (Gamalielsson et al. 2021).

Um solche Konflikten aus dem Weg zu gehen, empfiehlt sich die Verwendung eines quelloffenen und datenschutzkonformen Webanalysewerkzeugs wie beispielsweise *Matomo Analytics*. In der Studie, die dem Praxisbeitrag zur Nutzungsdatenanalyse (siehe auch Buckermann und Greving in diesem Band) zugrunde liegt, wurde diese Software verwendet. *Matomo* gilt nicht nur als datenschutzkonforme Open-Source-Alternative zu proprietärer Analyse-Software. Auch das *Data Ownership* ohne Ausnahme, das insbesondere bei Citizen-Science-Projekten zunehmend an Bedeutung gewinnt, ist eine der Kernfunktionalitäten von *Matomo*. Auch deswegen eignet sich diese Software als Instrument für die evaluative Wissenschaftskommunikationsforschung.

Die DSGVO und die dadurch begrenzte Verwendbarkeit von Analyse-Werkzeugen, die auf Browser-Cookies basieren, schränken die Anwendungsfelder von Nutzungsdatenanalysemethoden ein. Grundlegende Abrufzahlen sowie durch Webseiten-Besucher:innen verwendete Suchbegriffe sind nach wie vor problemlos zu erhalten; jedoch die Information darüber, ob ein:e Besucher:in ein Angebot mehrfach besucht hat, ist nicht mehr ohne dass Vorschalten eines sogenannten *Cookie-Banners* möglich, durch das das Einverständnis eine:r Besucher:in zur Erfassung dieser Daten eingeholt werden muss. Verzichtet man jedoch auf die Erfassung von Informationen, für deren Erhebung die Installation von Browser-Cookies notwendig ist, und anonymisiert andere personenbezogene Daten wie Besucher:innen-IP-Adressen, ist die Vorschaltung eines solchen Banners unter Umständen nicht nötig (Matomo 2021).

6 Fazit

Nutzungsdatenanalyse kann dabei unterstützen, mehr über die eigenen Wissenschaftskommunikationsprojekte zu erfahren, sie zu hinterfragen und sie zu verbessern. Sie hilft dabei, technische Fehlerquellen zu identifizieren und Barrierefreiheit zu gewährleisten. Ihre Werkzeuge können darüber hinaus vor allem für Fragestellungen, die auf das Engagement der Benutzer:innen in partizipativen Formaten abzielen, interessant sein. Dafür bietet sich häufig an,

Methoden der Nutzungsdatenanalyse mit anderen quantitativen oder qualitativen Methoden und Data-Scraping-Techniken zu kombinieren.

In der institutionellen Wissenschaftskommunikation ist es üblich, dass einzelne Institutionen mehrere Wissenschaftskommunikationsprojekte durchführen. In diesen Fällen empfiehlt es sich, mittels Web Analytics im Auge zu behalten, wie viele Besucher:innen durch welche Projekte der Wissenschaftskommunikation auf die Seite der Institution finden. So können z. B. Entscheidungen darüber, welche Ressourcen zukünftig auf welches Projekt verwendet werden, informierter getroffen werden. Hier ergibt sich jedoch zugleich eine wichtige Einschränkung: Kennzahlen der Nutzungsdatenanalyse allein lassen keine Aussagen über den quantitativen (Mehr-) Wert einzelner Maßnahmen oder gar ganzer Projekte zu. Ganz generell erscheint es sinnvoll, sich vor oder während der Einrichtung eines Wissenschaftskommunikationsprojekts einen Auszug der verfügbaren Analysedaten anzusehen, um sich mit den verschiedenen Metriken, die potenziell zur Verfügung stehen, vertrauter machen zu können. Es gilt daher basierend auf dem Gebot der Datensparsamkeit, nur diejenigen Daten, die für die Beantwortung konkreter Forschungs- oder Evaluationsfragen notwendig sind, zu erheben. Nach Formulierung konkreter Ziele und Forschungsfragen für das Projekt sollten überflüssige Metriken verworfen werden.

Trotz der Einführung der DSGVO und generell strengerer Richtlinien im Umgang mit Forschungs- wie persönlichen Daten können Nutzungsdatenanalyse-Methoden mit verhältnismäßig geringem Personal- oder Sachaufwand durchgeführt werden. Wie Nutzungsdatenanalyse datenschutzkonform und datensparsam durchgeführt werden kann, wird im Einzelnen bei Karg und Thomsen (2011) und Matomo (2021) beschrieben. Im Zweifel sollte das Vorgehen gemeinsam mit dem/der zuständigen Datenschutzbeauftragten koordiniert werden.

Für die evaluative Forschung an Wissenschaftskommunikationsprojekten gilt es, die Nutzungsdatenanalyse als Methodenset weiter zu elaborieren. Neben einer Schärfung der Instrumente bedarf es der weiteren Auseinandersetzung mit ihren Stärken und Schwächen, um methodischen, ethischen und praktischen Herausforderungen der Ansätze zu begegnen. Angesichts des rasanten Wachstums sowie des steten Wandels digitaler Kommunikationsformate sollten sich Praktiker:innen und Theoretiker:innen gleichermaßen mit der Nutzungsdatenanalyse befassen – um so auch zukünftig neue, digitale Wissenschaftskommunikation zu ihrem Betrachtungsgegenstand machen zu können.

Literatur

Batrinca B, Treleaven PC (2015) Social media analytics: a survey of techniques, tools and platforms. AI & Soc 30:89–116. https://doi.org/10.1007/s00146-014-0549-4

Bekavac I, Garbin Praničević D (2015) Web analytics tools and web metrics tools: an overview and comparative analysis. Croat Oper Res Rev 6 (2), 373–386. https://doi.org/10.17535/crorr.2015.0029

Boda K, Földes ÁM, Gulyás GG, Imre S (2012) User tracking on the web via cross-browser fingerprinting. In: Laud P (Hrsg) Information security technology for applications. NordSec 2011. Lecture notes in computer science 7161. Springer, Berlin. https://doi.org/10.1007/978-3-642-29615-4_4

Clifton B (2012) Advanced web metrics with google analytics, 3. Aufl. Sybex

Deutsche Forschungsgemeinschaft [DFG] (2019) Guidelines for safeguarding good research practice. Code of conduct. https://doi.org/10.5281/zenodo.3923602

Deutsche Forschungsgemeinschaft [DFG] (2021) Fachspezifische Empfehlungen zum Umgang mit Forschungsdaten. https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/empfehlungen/index.html. Zugegriffen: 30. Jan. 2022

Essig J, Watts M, Beck Dallaghan GL, Gilliland KO (2020) InstaHisto: utilizing Instagram as a medium for disseminating visual educational resources. Med Sci Educ 30(3):1035–1042. https://doi.org/10.1007/s40670-020-01010-2

Fagan JC (2013) The suitability of web analytics key performance indicators in the academic library environment. J Acad Librariansh 40(1):25–34. https://doi.org/10.1016/j.acalib.2013.06.005

Fritz W (2004) Internet-Marketing und Electronic Commerce. Grundlagen – Rahmenbedingungen – Instrumente. Gabler, Springer Fachmedien Wiesbaden GmbH, Wiesbaden. https://doi.org/10.1007/978-3-663-06001-7

Gamalielsson J, Lundell B, Butler S, Brax C, Persson T, Mattsson A, Gustavsson T, Feist J, Lönroth E (2021) Towards open government through open source software for web analytics: the case of Matomo. JeDEM EJ EDemocr Open Govt 13(2):133–153. https://doi.org/10.29379/jedem.v13i2.650

Guba A, Gebert O (1998) Online-Monitoring: Gewinnung und Verwertung von Online-Daten. In: Arbeitspapiere WI, Nr. 8/1998. Lehrstuhl für Allg BWL und Wirtschaftsinformatik (Hrsg) Johannes-Gutenberg-Universität, Mainz

Hallam-Baker PM, Behlendorf Brian (1996) Extended log file format. https://www.w3.org/TR/WD-logfile.html. Zugegriffen: 30. Jan. 2022

Karg M, Thomsen S (2011) Einsatz von Piwik bei der Reichweitenanalyse. DuD 35(489). https://doi.org/10.1007/s11623-011-0120-0

Kumar V, Ogunmola GA (2020) Web analytics for knowledge creation: a systematic review of tools, techniques, and practices. Int J Cyber Behav Psychol Learn 10(1):1–14. https://doi.org/10.4018/IJCBPL.2020010101

Krug S (2014) Don't make me think, revisited: a common sense approach to Web usability. New Riders Publishing, Thousand Oaks. https://dl.acm.org/doi/10.5555/1051204

McFadden C (2005) Optimizing the online business channel with web analytics. https://web.archive.org/web/20051221192230/ http://www.webanalyticsassociation.org:80/en/art/?9. Zugegriffen: 13. Febr. 2022

Matomo (2021) How do I use Matomo Analytics without consent or cookie banner? https://web.archive.org/web/20210824213215/ https://matomo.org/faq/new-to-piwik/how-do-i-use-matomo-analytics-without-consent-or-cookie-banner/. Zugegriffen: 13. Febr. 2022

Neuberger C, Weingart P, Fähnrich B, Fecher B, Schäfer MS, Schmid-Petri H, Wagner GG (2021) Der digitale Wandel der Wissenschaftskommunikation. In: Wissenschaftspolitik im Dialog: eine Schriftenreihe der Berlin-Brandenburgischen Akademie der Wissenschaften (Hrsg) Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

Podigee (2019) Unified podcast analytics whitepaper. https://web.archive.org/web/20220116121720/ https://docs.google.com/document/d/e/2PACX-1vQQcfYvUUsaR2UBPudmCpe5yNDs5HczB2ST9M6jEBwl3eHZyn5TcFX_HiEVSBE9y7yTBdZzF8HbRDdn/pub. Zugegriffen: 30. Jan. 2022

Priemer B (2004) Logfile-Analysen: Möglichkeiten und Grenzen ihrer Nutzung bei Untersuchungen zur Mensch-Maschine-Interaktion. MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung. Occasional Papers, S 1–23. https://doi.org/10.21240/mpaed/00/2004.06.02.X

Tandoc EC Jr, Thomas RJ (2015) The ethics of web analytics. Digit J 3(2):243–258. https://doi.org/10.1080/21670811.2014.909122

Welker M, Wünsch C (2010) Methoden der Online-Forschung. In: Schweiger W, Beck K (Hrsg) Handbuch Online-Kommunikation. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92437-3_20

Ziegler M (2014) Marktforschung. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung, S 183–193

Ziegler R, Fischer L (2020) Ziele von Wissenschaftskommunikation – Eine Analyse der strategischen Ziele relevanter Akteure für die institutionelle Wissenschaftskommunikation in Deutschland, 2014–2020. Wissenschaft im Dialog, Berlin

Armin Hempel ist wissenschaftlicher Mitarbeiter im Teilprojekt Öffentlichkeitsarbeit des DFG-SFBs 980 „Episteme in Bewegung“ an der Freien Universität Berlin. Er befasst sich mit digitalen Formaten, vorrangig mit Podcasts als Mittel der Wissenschaftskommunikation.



Praxisbeitrag: Nutzungsdatenanalyse digitaler Medien in der evaluativen Wissenschaftskommunikations-forschung am Beispiel eines Bürger-wissenschaftsprojekts

Till Bruckermann und Hannah Greving

Zusammenfassung

Durch die fortschreitende Digitalisierung finden verschiedene Formen der Wissenschaftskommunikation zunehmend online statt. Insbesondere in Bürgerwissenschaftsprojekten können Entwicklungen zu digitalen Formen der Wissenschaftskommunikation genutzt werden, um ihre Effektivität zu evaluieren: Die Analyse des Nutzungsverhaltens und der durchgeführten Aktivitäten der Teilnehmenden in der Online-Umgebung des Projekts kann Aufschluss darüber geben, was Teilnehmende genau machen und womit sie sich beschäftigen. Diese Erkenntnis kann wiederum die Kommunikation mit Teilnehmenden verbessern. Am Beispiel eines Bürgerwissenschaftsprojekts zu Wildtieren wird die Analyse von Nutzungsdaten in der Wissenschafts-kommunikation im Hinblick auf die Häufigkeit und Art der Beteiligung in Projektaktivitäten erläutert. Es wird deutlich, dass Beteiligungsmuster der Teilnehmenden von den intendierten Aktivitäten abweichen. Implikationen für das Lernen aus Bürgerwissenschaftsprojekten werden diskutiert.

T. Bruckermann (✉)
Institut für Erziehungswissenschaft, Leibniz Universität Hannover, Hannover, Deutschland
E-Mail: till.bruckermann@iew.uni-hannover.de

H. Greving
Arbeitsgruppe Wissenskonstruktion, Leibniz-Institut für Wissensmedien, Tübingen, Tübingen, Deutschland
E-Mail: h.greving@iwm-tuebingen.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_11

Durch die fortschreitende Digitalisierung finden verschiedene Formen der Wissenschaftskommunikation zunehmend online über digitale Medien statt (Neuberger et al. 2021). Diese Entwicklungen können genutzt werden, um die Effektivität von Wissenschaftskommunikation zu evaluieren: Die Analyse des Nutzungsverhaltens und der durchgeführten Aktivitäten der Nutzer:innen in einer Online-Umgebung kann Aufschluss darüber geben, was Nutzer:innen genau machen und womit sie sich beschäftigen. Diese Erkenntnis kann wiederum zur Verbesserung von Wissenschaftskommunikation beitragen. Der vorliegende Beitrag erläutert die Analyse von Nutzungsdaten in der Wissenschaftskommunikation am Beispiel eines Bürgerwissenschaftsprojekts zu Wildtieren.

1 Wissenschaftskommunikation in Bürgerwissenschaftsprojekten

In Bürgerwissenschaftsprojekten arbeiten engagierte Bürger:innen zusammen mit professionell arbeitenden Wissenschaftler:innen in wissenschaftlichen Projekten (Heigl et al. 2019). Häufig sammeln oder verarbeiten Bürger:innen in solchen Projekten vor allem Daten für Wissenschaftler:innen. Es gibt allerdings auch Ansätze, in denen Bürger:innen und Wissenschaftler:innen gemeinsam wissenschaftliches Wissen generieren und auch jeweils ihr eigenes Wissen über (Natur-) Wissenschaften weiterentwickeln (Bonney et al. 2016). Damit folgen diese Projekte einem Paradigmenwechsel in der Wissenschaftskommunikation von einer unidirektionalen Kommunikation über Wissenschaft, die sich gemäß dem Defizitmodell ausschließlich von Wissenschaftler:innen an das Laienpublikum richtet, zu einer partizipativen Form von Kommunikation über Wissenschaft, bei welcher der multidirektionale Austausch im Vordergrund steht (Trench 2008). Wenn sie einen multidirektionalen Austausch von Wissen zwischen Bürger:innen und Wissenschaftler:innen ermöglichen, können Bürgerwissenschaftsprojekte als partizipative Form von Wissenschaftskommunikation verstanden werden (Wagenknecht et al. 2021).

Der Mehrwert dieser partizipativen Form wurde aus einer normativen Perspektive auch als „Leiter“ beschrieben (Arnstein 1969), auf der Bürger:innen zu weitestgehender Beteiligung aufsteigen sollten. In der Folge wurden Modelle zur Beteiligung von Bürger:innen an wissenschaftlichen Projekten aufgestellt, die zwischen den Angeboten zur Beteiligung und der tatsächlichen Beteiligung unterscheiden (Shirk et al. 2012; für eine Erweiterung siehe Bruckermann et al. 2020). Des Weiteren wurden die Projekte je nach Möglichkeit zur Beteiligung der Bürger:innen in verschiedene Projektkategorien eingeteilt (vgl. Haklay 2018,

für eine Übersicht). Beispielsweise haben Bürger:innen in *contributory projects* alleinig die Möglichkeit zur Datensammlung, während sie in *collaborative projects* die Möglichkeit zur Datensammlung und Datenaufbereitung sowie zur Datenanalyse haben (Shirk et al. 2012). Dementsprechend kann die Beteiligung von Bürger:innen anhand der Nutzung dieser Möglichkeiten zur Beteiligung in den unterschiedlichen Aktivitäten *qualifiziert* werden (für eine Typologie siehe Wiggins und Crowston 2012). Bei dieser qualitativen Beschreibung der Beteiligung wird angenommen, dass Bürger:innen umso mehr aus einem Bürgerwissenschaftsprojekt mitnehmen, wenn sie in kognitiv anspruchsvollere Aufgaben eingebunden werden (Shirk et al. 2012). Neben der qualitativen Beschreibung der Beteiligung kann die Beteiligung auch auf *quantitative* Art und Weise beschrieben werden. Eine solche Beschreibung folgt der Annahme, dass Bürger:innen desto mehr aus den Projekten lernen, je häufiger sie sich an einem Projekt beteiligen.

2 Nutzungsdatenanalyse digitaler Beteiligung in Bürgerwissenschaftsprojekten

Seit dem Christmas Bird Count zu Anfang des 20. Jahrhunderts haben sich Bürgerwissenschaftsprojekte durch die Verfügbarkeit digitaler Technologien verändert (Preece 2016). Nicht nur wissenschaftliche Daten können in größerem Maße gesammelt (z. B. iSpot; Silvertown et al. 2015), auch Projekte können einfacher über Internetplattformen initiiert werden (z. B. nQuire-it; Aristeidou et al. 2017). Des Weiteren können Bürger:innen räumlich unabhängig an Bürgerwissenschaftsprojekten auf Internetplattformen teilnehmen (Preece 2016) und haben Werkzeuge zur Verfügung, mit denen sie Daten auswerten und die Ergebnisse diskutieren können (Bonney et al. 2014). Durch die Online-Beteiligung von Bürger:innen können deren Nutzungsdaten auf der Internetplattform (nach Zustimmung) erfasst und analysiert werden. Beispielsweise werteten einige Projekte aus, wie intensiv die Kommunikation und Kollaboration zwischen Bürger:innen und Wissenschaftler:innen ausfiel (z. B. Cox et al. 2015), ob sich gewisse Gruppen von Bürger:innen aus Beteiligungsprofilen identifizieren ließen (z. B. Aristeidou et al. 2017) und wie die Beteiligung mit bestimmten Eigenschaften der Bürger:innen zusammenhing (vgl. Aristeidou und Herodotou 2020, für eine Übersicht).

Die Auswertung von Nutzungsdaten folgt den Zielen des Bürgerwissenschaftsprojekts. Vorrangig an wissenschaftlichen Zielen orientierte Projekte, die Kommunikation über Wissenschaft nachrangig betrachten, nehmen in der

Nutzungsdatenanalyse häufig eine quantitative Sicht auf die Beteiligung von Bürger:innen ein. Beispielsweise erfassen solche Projekte die Beteiligung als Zeiteinheiten, in denen Bürger:innen in einem Projekt mitarbeiten, oder als Produkte, welche Bürger:innen im Projekt erarbeiten (d. h. Anzahl der Tage im Projekt und Anzahl der Klassifikationen bezogen auf z. B. Galaxien oder Pinguine; Masters et al. 2016). Wenn Bürgerwissenschaftsprojekte neben ihren wissenschaftlichen Zielen vorrangig den Dialog und die Partizipation von Bürger:innen verfolgen, indem sie Angebote zur Beteiligung an Aktivitäten über die Datensammlung hinaus machen, sollten diese Projekte neben der Quantität auch die Qualität der Beteiligung in den Blick nehmen. Aus qualitativer Sicht werden in Nutzungsdatenanalysen die verschiedenen Aktivitäten berücksichtigt, an denen sich Bürger:innen beteiligen können.

In Nutzungsdatenanalysen von Online-Bürgerwissenschaftsprojekten können die Häufigkeit (Quantität) und die Art (Qualität) der Beteiligung außerdem durch eine Unterscheidung von aktiver und passiver Beteiligung abgebildet werden (vgl. Malinen 2015, für einen Überblick). Bürger:innen, die sich aktiv beteiligen, tragen zum Projektziel bei, indem sie Daten beisteuern und auswerten sowie im Forum diskutieren. Passiv beteiligte Bürger:innen zeichnen sich hingegen durch eine beobachtende Haltung ohne eigene Beiträge auf der Internetplattform aus, was auch als *Stöbern* bezeichnet wird (Malinen 2015). Nutzungsdatenanalysen sollten nicht nur passiv Beteiligte identifizieren, sondern auch bei der Suche nach Gründen für die passive Beteiligung helfen, sodass passiv Beteiligte zu einer aktiven Beteiligung motiviert werden könnten (Eveleigh et al. 2014).

3 Nutzungsdatenanalyse am Beispiel der Beteiligung im Projekt Wildtierforscher in Berlin

Im folgenden Anwendungsbeispiel wurden die Nutzungsdaten in einem Bürgerwissenschaftsprojekt zu Wildtieren von den teilnehmenden Bürger:innen erfasst und analysiert. Dieses Projekt wird ebenfalls aufgegriffen im Beitrag zu experimentellen Herangehensweisen (siehe Greving et al. in diesem Band). Das Ziel des Bürgerwissenschaftsprojekts *Wildtierforscher in Berlin* war es aus wissenschaftlicher Sicht, die Ökologie und das Vorkommen von in Berlin lebenden terrestrischen Säugetieren zu untersuchen. Außerdem beschäftige sich das Projekt mit der Frage, was Bürger:innen durch die Teilnahme an diesem Bürgerwissenschaftsprojekt lernen (Bruckermann et al. 2021). Der Annahme folgend, dass Teilnehmende mehr aus Bürgerwissenschaftsprojekten mitnehmen, wenn sie sich an möglichst vielen Aktivitäten des Forschungsprozesses

beteiligen (Bonney et al. 2009), bot dieses Projekt nicht nur die Möglichkeit, dass Teilnehmende Daten sammeln, sondern darüber hinaus Daten auch statistisch auswerten und ihre Ergebnisse mit anderen Teilnehmenden und Wissenschaftler:innen diskutieren konnten.

Durch die Nutzungsdatenanalyse sollte die Forschungsfrage geklärt werden, ob sich die Teilnehmenden an der Datenauswertung in einem ähnlichen Umfang beteiligen wie an der Datensammlung. Auch wenn davon ausgegangen wird, dass Teilnehmende mehr lernen und somit beispielsweise mehr Wissen mitnehmen, wenn sie sich an anspruchsvolleren Tätigkeiten im Projekt beteiligen (Bonney et al. 2009), sind nicht alle Teilnehmenden auch motiviert, sich in anspruchsvollere Tätigkeiten wie der Datenanalyse einzubringen (Phillips et al. 2019). Der durch Befragungen zur Motivation von Teilnehmenden festgestellte Befund vorheriger Studien (Phillips et al. 2019) wurde bisher noch nicht durch Verhaltensdaten untermauert, die klar zeigen, dass Teilnehmende sich nicht an der Datenanalyse beteiligen, wenn ihnen die Möglichkeit angeboten wird. Die Nutzungsdatenanalyse eignet sich, um das Verhalten von Teilnehmenden im Bürgerwissenschaftsprojekt zu beschreiben, da sowohl die Häufigkeit (Quantität) als auch die Art der Aktivitäten im Projekt (Qualität) beschrieben werden kann.

Das Bürgerwissenschaftsprojekt umfasste insgesamt fünf Durchgänge von saisonalen Feldphasen (im Frühjahr und Herbst) zwischen 2018 und 2020, die jeweils in etwa zwei Monate dauerten. Bis zu 200 Bürger:innen aus Berlin beteiligten sich pro Feldphase an dem Projekt. Für die Zeit ihrer Teilnahme erhielten sie leihweise eine sogenannte Kamerafalle, die in der Lage ist, tagsüber und auch nachts Fotos von Wildtieren aufzunehmen. Während der Datensammlung stellten die Teilnehmenden diese Kamerafallen im eigenen Garten auf, um zu überprüfen, wie häufig in Berlin lebende terrestrische Säugetierarten in einem gewissen Zeitraum gesichtet werden können.

Alle weiteren Aktivitäten im Projekt fanden online auf einer eigens für dieses Projekt entwickelten Internetplattform statt, die einen nur für Teilnehmende zugänglichen Login-Bereich hatte. Während der Datensammlung mit der Kamerafalle konnten die Teilnehmenden auf der Plattform ein Tutorial absolvieren, das ihnen erklärte, wie die Wildtiere auf den Fotos der Kamerafalle zu bestimmen sind. Außerdem konnten sie die Fotos ihrer Kamerafalle auf die Plattform hochladen und die Tierarten auf den eigenen Fotos sowie auf Fotos anderer Teilnehmenden bestimmen, um deren jeweilige Bestimmung zu validieren. Nach der Datensammlung konnten die Teilnehmenden während der Datenauswertung die eigenen erfassten Daten oder den gesamten Datensatz aller teilnehmenden Bürger:innen auswerten und beispielsweise analysieren, welche Umweltvariablen (wie z. B. Versiegelungsgrad und Baumbedeckung)

das Vorkommen verschiedener Säugetierarten beeinflussen. Weiterhin konnten sie ihre Ergebnisse und Fragen mit anderen Teilnehmenden und den beteiligten Wissenschaftler:innen im Forum diskutieren. Während des gesamten Projekts konnten sich die Teilnehmenden über die in Berlin vorkommenden Wildtiere sowie über die Stadt als Lebensraum für Wildtiere informieren.

4 Erfassung und Aufbereitung der Nutzungsdaten zur Beteiligung

Um die Beteiligung der Bürger:innen zu evaluieren, wurden Nutzungsdaten auf der Internetplattform des Projekts erhoben. Die Nutzungsdaten wurden mit der freien Open-Source-Anwendung Matomo (v3.9.1) erhoben. Mit dieser Anwendung wurde für jede teilnehmende Person genau erfasst, an welchen Tagen die Person eingeloggt war, wie lange sie dann auf welcher Seite der Plattform war und ob sie gegebenenfalls zusätzlich noch eine Aktion durchgeführt hat. Anhand dieser erfassten Daten wurden die Tage anhand der aus bisheriger Forschung bekannten Unterscheidung auf Internetplattformen in Tage aktiver und passiver Beteiligung eingeteilt (Malinen 2015). An sogenannten *aktiven Tagen* führten die Teilnehmenden mindestens eine Aktion auf der Plattform aus. Das heißt sie luden beispielsweise Fotos hoch, absolvierten das Tutorial, identifizierten oder validierten die Tierarten auf den Fotos, analysierten Daten oder posteten eine Frage, einen Kommentar oder ihre Ergebnisse im Forum. An *passiven Tagen* führten die Teilnehmenden keine dieser Aktionen durch, sondern durchstöberten stattdessen die Inhalte der Plattform. Das heißt, sie schauten sich ihre eigenen Fotos oder die der anderen Teilnehmenden an, lasen die Beiträge im Forum oder beschäftigten sich mit den Informationen zu den Wildtieren und deren städtischem Lebensraum auf der Plattform. Summiert man die Tage ohne Login und die passiven Tage zwischen zwei aktiven Tagen, ergibt sich die Anzahl der *Tage zwischen zwei aktiven Tagen*. Die Differenz zwischen dem ersten und dem letzten Tag des Logins definierte die *gesamten Tage*, an denen eine teilnehmende Person mit dem Projekt verbunden war. Abschließend ergaben die Tage zwischen dem ersten Tag des Logins und dem Tag, an dem das Projekt endete, die *potenziellen Tage,* an denen eine teilnehmende Person mit dem Projekt verbunden sein konnte.

Für eine *quantitative Betrachtung* der Beteiligung wurden nun anhand der Einteilung in passive sowie aktive Tage und auf Basis von früheren Studien (Aristeidou et al. 2017; Ponciano und Brasileiro 2014) aus den Nutzungsdaten Kennzahlen berechnet. Um diese Kennzahlen für die Teilnehmenden berechnen zu können, mussten diese mindestens zwei Tage lang auf der Plattform aktiv

gewesen sein. Die Kennzahlen ergaben sich wie folgt. Der *Aktivitätsquotient* beschrieb das Verhältnis von aktiven Tagen zu den gesamten Tagen, die eine teilnehmende Person mit dem Projekt verbunden war. Der *Stöberquotient* beschrieb das Verhältnis zwischen den passiven Tagen und den potenziellen Tagen, an denen eine teilnehmende Person mit dem Projekt verbunden sein konnte. Die an aktiven Tagen *aufgewendete Zeit* beschrieb die Anzahl der Stunden, die eine teilnehmende Person an aktiven Tagen für ihren Beitrag aufgebracht hatte. Die *Regelmäßigkeit* der aktiven Beteiligung wurde bestimmt, indem die Standardabweichung der Anzahl der Tage zwischen allen aufeinanderfolgenden aktiven Tagen berechnet wurde. Schlussendlich ergab sich die *relative Dauer* der Beteiligung im Projekt aus dem Verhältnis zwischen den Gesamttagen und den potenziellen Tagen einer teilnehmenden Person im Projekt. Zur Beantwortung der Fragestellung im Projekt war zusätzlich eine *qualitative Betrachtung* der Beteiligung notwendig, indem die in den Nutzungsdaten erfassten Aktivitäten auf der Internetplattform entsprechend der Typologie (Wiggins und Crowston 2012) entweder der Datensammlung (Bilder hochladen, bestimmen und validieren; Tutorial zur Bildbestimmung ansehen) oder der Datenauswertung (statistische Auswertungen durchführen; Ergebnisse im Forum hochladen und diskutieren) zugeordnet wurden. Deshalb wurden die quantitativen Kennzahlen auch qualitativ für jede teilnehmende Person getrennt für die Datensammlung und die Datenauswertung berechnet.

5 Erkenntnisse zur Beteiligung in Datensammlung und -auswertung

Bei der Analyse zeigten sich anhand der berechneten Kennzahlen abweichende Beteiligungsmuster. Der mittlere *Aktivitätsquotient* war während der Datensammlung höher als während der Datenauswertung. Gleichzeitig war der *Stöberquotient* im Durchschnitt während der Datensammlung niedriger als während der Datenauswertung. Ein ähnliches Muster zeigte sich für die an aktiven Tagen *aufgewendete Zeit* insofern, dass Teilnehmende während der Datensammlung im Mittel mehr Zeit mit Aktivitäten verbrachten als während der Datenauswertung. Außerdem beteiligten sich die Teilnehmenden an der Datensammlung *regelmäßiger* als an der Datenauswertung und von der *relativen* Dauer her blieben sie bei der Datensammlung auch länger dabei als bei der Datenauswertung. Zusammengenommen deuteten die Kennzahlen darauf hin, dass Teilnehmende sich an der Datensammlung aktiver beteiligten, mehr Zeit aufwandten und sich über einen längeren Zeitraum einbrachten als an der Datenauswertung.

Diese Befunde stützen die Ergebnisse vorheriger Forschung zu den Motiven von Projektteilnehmenden, die sich eher an der Sammlung von Daten beteiligen wollten (Phillips et al. 2019), indem sie diese Ergebnisse aus Fragebogendaten um konkrete Verhaltensdaten von Teilnehmenden aus einer Nutzungsdatenanalyse erweitern (Bruckermann et al. 2022).

Weitere Vorteile einer qualitativen Betrachtung von Nutzungsdaten gegenüber einer rein quantitativen Betrachtung zeigen sich in folgendem Vergleich mit einem anderen Bürgerwissenschaftsprojekt. Im Vergleich zum Projekt *Weather-it*, in dem die Teilnehmenden Wetterdaten sammelten und teilten (Aristeidou et al. 2017), zeigte sich, dass Teilnehmende im Projekt *Wildtierforscher in Berlin* im Mittel weniger aktiv waren (d. h. einen geringeren Aktivitätsquotient hatten), aber nur, wenn die Feldphasen nicht nach Datensammlung und Datenauswertung differenziert betrachtet wurden. Wenn allerdings die Kennzahlen während der Datensammlung und Datenauswertung getrennt betrachtet wurden, war während der Datensammlung im Projekt *Wildtierforscher in Berlin* der Aktivitätsquotient vergleichbar zum Aktivitätsquotienten im Projekt *Weather-it* (Bruckermann et al. 2022). Dieser Unterschied verdeutlicht, dass eine rein quantitative Sicht auf Nutzungsdaten nicht ausreichend ist und die bloße Evaluation anhand von Tagen mit Seitenaufrufen und Verweildauern nicht ausreicht. Vielmehr müssen Nutzungsdaten qualitativ, beispielsweise entlang einer Typologie (Wiggins und Crowston 2012), differenziert werden, insbesondere wenn vorher Annahmen zur Beteiligung getroffen wurden (Bonney et al. 2009). In der Praxis kann sich eine Nutzungsdatenanalyse zur Projektevaluation in der Wissenschaftskommunikation einerseits auf quantitative Kennzahlen, wie beispielsweise die aufgewendete Zeit, stützen, die relativ vergleichbar zwischen unterschiedlichen Projekten ist (siehe Vergleich mit *Weather-it*). Andererseits sollten unbedingt auch die Eigenschaften der jeweiligen Projekte in der Nutzungsdatenanalyse für eine qualitative Betrachtung berücksichtigt werden. Daher empfiehlt sich für die Praxis, die Eigenschaften des eigenen Projektes in der Analyse zu berücksichtigen und zu anderen Projekten anhand einer Typologie (z. B. Typologie von Projektaktivitäten in Bürgerwissenschaftsprojekten; Wiggins und Crowston 2012) in ein Verhältnis zu setzen.

Die Nutzungsdatenanalyse bringt allerdings auch Limitationen mit sich. Diese liegen in der Beschränkung auf erfassbares Verhalten und der Fülle und deskriptiven Natur verfügbarer Daten. Insbesondere das Verhalten passiv beteiligter Personen ist schwierig zu erfassen, da es nicht durch Beiträge auf der Internetplattform sichtbar wird (Malinen 2015). Andererseits werden in Nutzungsdaten eine Fülle an Informationen über das Verhalten der Teilnehmenden auf der Internetplattform erfasst, die zunächst aufbereitet und dann interpretiert werden müssen. Um in Online-Bürgerwissenschaftsprojekten das

Verhalten von Teilnehmenden aus Nutzungsdaten im Speziellen abzubilden (Aristeidou et al. 2017), kann beispielsweise auf die theoretischen Rahmungen vorheriger Arbeiten zu Online-Communities im Internet im Allgemeinen zurückgegriffen werden (Malinen 2015). Die deskriptive Natur von Nutzungsdaten bedingt, dass diese zunächst einmal nur das Verhalten der Teilnehmenden auf der Interplattform abbilden. Sie erklären aber noch nicht, warum Teilnehmende während der Datensammlung aktiver waren als während der Datenauswertung. Deshalb scheint eine Kombination von Nutzungsdaten mit weiteren, beispielsweise aus Fragebögen gewonnen Daten, der Evaluation zuträglich, wenn Schlussfolgerungen gezogen werden sollen.

6 Fazit

Aus Nutzungsdatenanalysen gewonnene Erkenntnisse über das Verhalten von Teilnehmenden auf der Internetplattform ermöglichen Rückschlüsse auf die Nutzung von Angeboten in Bürgerwissenschaftsprojekten (MODEL-CS; Bruckermann et al. 2020, 2022). Für die Praxis ist eine Unterscheidung zwischen den Angeboten einer Internetplattform und ihrer Nutzung, zu der die Nutzungsdatenanalyse Erkenntnisse liefern kann, zentral, wenn es um die Erklärung geht, weshalb gewisse Formate der Wissenschaftskommunikation Effekte beispielsweise auf das Wissen haben. Aus den Ergebnissen ergeben sich deshalb weitere Fragen für die Projektevaluation, die sowohl die Förderung von Beteiligung als auch Lerneffekte aus der Beteiligung in Bürgerwissenschaftsprojekten betreffen. Bei Bürgerwissenschaftsprojekten, die als *collaborative projects* bezeichnet werden, stellt sich die Frage, ob Lerneffekte vor allem auf eine Beteiligung an anspruchsvolleren Tätigkeiten wie der Datenauswertung zurückgeführt werden können, wenn sich Teilnehmende, wie beobachtet, vor allem an der Datensammlung aktiv beteiligen. Um diese Frage klären zu können, wäre eine Kombination mit Daten aus Fragebogenerhebungen notwendig (siehe auch Böhmert und Abacioglu in diesem Band). Auch wenn es aus Fall- und Korrelationsstudien Hinweise gibt, dass Qualität und Quantität der Beteiligung mit Lerneffekten im Zusammenhang stehen könnten (Masters et al. 2016; Shirk et al. 2012), fand eine experimentelle Feldstudie (siehe auch Greving et al. in diesem Band) keine Unterschiede zwischen Teilnehmenden, die nur Daten gesammelt hatten, und solchen, die Daten gesammelt und auch ausgewertet hatten (Greving et al. 2022). Weitere Forschung sollte daher die unterschiedlichen Arten der Beteiligung genauer betrachten. In der Praxis empfiehlt sich deshalb die Nutzungsdatenanalyse mit weiteren Datenquellen aus Befragungen zu kombinieren, wenn in der Wissenschaftskommunikation evaluiert werden soll,

weshalb gewisse Angebote genutzt werden und welche Effekte diese Angebote beispielsweise auf das Wissen der Teilnehmenden haben. Außerdem zeigen die Befunde, dass es vermutlich nicht reicht, den Teilnehmenden nur die Möglichkeit zur Beteiligung in verschiedenen Phasen des Forschungsprozesses anzubieten, da sie hiervon unterschiedlich Gebrauch machen. Daraus ergibt sich die Frage, wie eine Beteiligung bei der Auswertung von Daten gefördert werden kann. In der Praxis kann die Nutzungsdatenanalyse also Hinweise geben, welche Angebote in der Wissenschaftskommunikation nicht wie intendiert genutzt werden. Weitere Ergebnisse aus Laborstudien im Projekt *Wildtierforscher in Berlin* deuten darauf hin, dass insbesondere das Gefühl, die Daten zu besitzen, und die wahrgenommene Rolle im Projekt die aktive Beteiligung an der Datenauswertung fördern könnten (Greving et al. 2020). Zudem benötigen die Teilnehmenden möglicherweise auch eine stärkere Begleitung und Anleitung, um sich aktiver an der Datenauswertung zu beteiligen. Um also die Motive hinter dem gezeigten Verhalten zu erschließen oder Rückschlüsse aus dem Verhalten auf die Wirkung eines Angebots ziehen zu können, wäre es sinnvoll, in der Zukunft Nutzungsdaten mit weiteren Informationen über die Teilnehmenden aus Befragungen oder Beobachtungen zur Evaluation von Formaten der Wissenschaftskommunikation zu kombinieren.

Literatur

Aristeidou M, Herodotou C (2020) Online citizen science: a systematic review of effects on learning and scientific literacy. Citiz Sci Theory Pract 5(1):1–12. https://doi.org/10.5334/cstp.224

Aristeidou M, Scanlon E, Sharples M (2017) Profiles of engagement in online communities of citizen science participation. Comput Hum Behav 74:246–256. https://doi.org/10.1016/j.chb.2017.04.044

Arnstein SR (1969) A ladder of citizen participation. J Am Inst Plan 35(4):216–224. https://doi.org/10.1080/01944366908977225

Bonney RE, Cooper CB, Dickinson J, Kellling S, Phillips T, Rosenberg KV, Shirk J (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59(11):977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bonney RE, Shirk JL, Phillips TB, Wiggins A, Ballard HL, Miller-Rushing AJ, Parrish JK (2014) Citizen science. Next steps for citizen science. Science 343(6178):1436–1437. https://doi.org/10.1126/science.1251554

Bonney RE, Phillips TB, Ballard HL, Enck JW (2016) Can citizen science enhance public understanding of science? Public Underst Sci 25(1):2–16. https://doi.org/10.1177/0963662515607406

Bruckermann T, Lorke J, Rafolt S, Scheuch M, Aristeidou M, Ballard H, Bardy-Durchhalter M, Carli E, Herodotou C, Kelemen-Finann J, Robinson L, Swanson R, Winter S, Kapelari S (2020) Learning opportunities and outcomes in citizen science: a

heuristic model for design and evaluation. In: Levrini O, Tasquier G (Hrsg) Electronic proceedings of the ESERA 2019 conference. The beauty and pleasure of understanding: engaging with contemporary challenges through science education. University of Bologna, Bologna, S 889–898

Bruckermann T, Greving H, Schumann A, Stillfried M, Börner K, Kimmig SE, Hagen R, Brandt M, Harms U (2021) To know about science is to love it? Unraveling cause–effect relationships between knowledge and attitudes toward science in citizen science on urban wildlife ecology. J Res Sci Teach 58(8):1179–1202. https://doi.org/10.1002/tea.21697

Bruckermann T, Greving H, Stillfried M, Schumann A, Brandt M, Harms U (2022) I'm fine with collecting data: Engagement profiles differ depending on scientific activities in an online community of a citizen science project PLoS ONE 17:e0275785. https://doi.org/10.1371/journal.pone.0275785

Cox J, Oh YE, Simmons B, Lintott C, Masters K, Greenhill A, Graham G, Holmes K (2015) Defining and measuring success in online citizen science: a case study of zooniverse projects. Comput Sci Eng 17(4):28–41. https://doi.org/10.1109/MCSE.2015.65

Eveleigh A, Jennett C, Blandford A, Brohan P, Cox AL (2014) Designing for dabblers and deterring drop-outs in citizen science. In: Jones M, Palanque P, Schmidt A, Grossman T (Hrsg) CHI'14: proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, S 2985–2994

Greving H, Bruckermann T, Kimmerle J (2020) This is my project! The influence of involvement on psychological ownership and wildlife conservation: the influence of involvement on psychological ownership and wildlife conservation. Curr Res Ecol Soc Psychol 1:100001. https://doi.org/10.1016/j.cresp.2020.100001

Greving H*, Bruckermann T*, Schumann A, Straka TM, Lewanzik D, Voigt-Heucke SL, Marggraf L, Lorenz J, Brandt M, Voigt C, Harms U, Kimmerle J (2022) Improving attitudes and knowledge in a citizen science project on urban bat ecology. Ecol Soc 27(2):Article 24. *shared first-authorship. https://doi.org/10.5751/es-13272-270224

Haklay M (2018) Participatory citizen science. In: Hecker S et al (Hrsg) Citizen science: innovation in open science, society and policy. UCL Press, London, S 52–62

Heigl F, Kieslinger B, Paul KT, Uhlik J, Dörler D (2019) Opinion: toward an international definition of citizen science. Proc Natl Acad Sci USA 116(17):8089–8092. https://doi.org/10.1073/pnas.1903393116

Malinen S (2015) Understanding user participation in online communities: a systematic literature review of empirical studies. Comput Hum Behav 46:228–238. https://doi.org/10.1016/j.chb.2015.01.004

Masters K, Oh EY, Cox J, Simmons B, Lintott C, Graham G, Greenhill A, Holmes K (2016) Science learning via participation in online citizen science. J Sci Commun 15(03):A07. https://doi.org/10.22323/2.15030207

Neuberger C, Weingart P, Fähnrich B, Fecher B, Schäfer MS, Schmid-Petri H, Wagner GG (2021) Der digitale Wandel der Wissenschaftskommunikation. Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

Phillips TB, Ballard HL, Lewenstein BV, Bonney R (2019) Engagement in science through citizen science: moving beyond data collection. Sci Educ 103(3):665–690. https://doi.org/10.1002/sce.21501

Ponciano L, Brasileiro F (2014) Finding volunteers' engagement profiles in human computation for citizen science projects. Hum Comp 1(2):247–266. https://doi.org/10.15346/hc.v1i2.12

Preece J (2016) Citizen science: new research challenges for human–computer interaction. Int J Hum-Comp Interact 32(8):585–612. https://doi.org/10.1080/10447318.2016.1194153

Shirk JL, Ballard HL, Wilderman CC, Phillips T, Wiggins A, Jordan R, McCallie E, Minarchek M, Lewenstein BV, Krasny ME, Bonney R (2012) Public participation in scientific research: a framework for deliberate design. Ecol Soc 17(2):29. https://doi.org/10.5751/ES-04705-170229

Silvertown J, Harvey M, Greenwood R, Dodd M, Rosewell J, Rebelo T, Ansine J, McConway K (2015) Crowdsourcing the identification of organisms: a case-study of iSpot. ZooKeys 480:125–146. https://doi.org/10.3897/zookeys.480.8803

Trench B (2008) Towards an analytical framework of science communication models. In: Cheng D et al (Hrsg) Communicating science in social contexts: new models, new practices. Springer, Dordrecht, S 119–135

Wagenknecht K, Woods T, Nold C, Rüfenacht S, Voigt-Heucke S, Caplan A, Hecker S, Vohland K (2021) A question of dialogue? Reflections on how citizen science can enhance communication between science and society. J Sci Commun 20(03):A13. https://doi.org/10.22323/2.20030213

Wiggins A, Crowston K (2012) Goals and tasks: two typologies of citizen science projects. In: Sprague RH (Hrsg) 2012 45th Hawaii international conference on system sciences, IEEE computer society. IEEE, Piscataway, S 3426–3435

Till Bruckermann ist Universitätsprofessor an der Leibniz Universität Hannover. Er forscht zu informellem Lernen in Bürgerwissenschaftsprojekten und insbesondere zur Entwicklung eines Wissenschaftsverständnisses.

Hannah Greving ist wissenschaftliche Mitarbeiterin am Leibniz-Institut für Wissensmedien in Tübingen. Sie beschäftigt sich mit dem Einfluss von Bürgerwissenschaftsprojekten auf die Teilnehmenden und effektiver Wissenschaftskommunikation.



Grundlagenbeitrag: Physiologische Messungen im Kontext der Evaluation von Wissenschaftskommunikation

Philipp Niemann und Yannic Scheuermann

Zusammenfassung

Der Grundlagenbeitrag zu physiologischen Messungen behandelt ausgewählte, praxistaugliche Verfahren aus einem breiten Methodenspektrum, die sich für wissenschaftskommunikative Evaluationsvorhaben eignen. Vorgestellt werden diese gegliedert nach Erregungs-, Emotions-, Aufmerksamkeits- und Bewertungsindikatoren. Besonderes Augenmerk liegt auf der Methode der Blickaufzeichnung sowie auf der Real-Time-Response-Messung, deren Anwendung in der Praxis der Wissenschaftskommunikation besonders naheliegend erscheinen. Im Vergleich zu anderen Evaluationsmethoden wird deutlich, dass der Einsatz physiologischer Messungen insgesamt sowohl in Konzeption, Durchführung und Auswertung komplex ist, jedoch ein tiefgehendes Verständnis für Rezeptionsprozesse bei wissenschaftskommunikativen Artefakten schaffen kann.

P. Niemann (✉) · Y. Scheuermann
Nationales Institut für Wissenschaftskommunikation gGmbH, Heidelberg, Deutschland
E-Mail: niemann@nawik.de

Y. Scheuermann
E-Mail: scheuermann@nawik.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,* https://doi.org/10.1007/978-3-658-39582-7_12

1 Einleitung

Physiologische Messungen „dienen der objektiven Erfassung und Quantifizierung bestimmter Merkmale physiologischer Prozesse in unterschiedlichen Organsystemen des Körpers mittels entsprechender Messgeräte“ (Döring und Bortz 2016, S. 501). Sie sind für die Rezeptions- und Wirkungsforschung und gerade auch für evaluatorische Untersuchungen deshalb von Interesse, weil sie unterschiedliche Indikatoren für menschliches Erleben bzw. Verhalten liefern können. Gegenüber anderen Methoden, etwa Selbstauskünften mittels Befragungen oder Tests, bieten physiologische Messungen einige Vorteile: Es ist deutlich schwieriger, sie bewusst zu manipulieren, Gedächtnisfehler spielen keine Rolle, es werden Phänomene zugänglich, die nicht bewusst wahrgenommen werden oder nicht präzise verbalisierbar sind (z. B. die Lidschlagfrequenz), und Rezeptionsgeschehen kann sehr genau im Zeitverlauf erfasst werden (z. B. mittels Blickbewegungserfasssung) – um nur die zentralen Vorzüge zu nennen (vgl. Döring und Bortz 2016, S. 502). Nachteile physiologischer Messungen sind beispielsweise der notwendige Zugriff auf kostspielige Spezialgeräte, die Notwendigkeit von hoher Fachexpertise zur Durchführung und Auswertung der Untersuchungen oder die Belastungen/Beeinträchtigungen von Proband:innen durch die Messgeräte inkl. der Reaktivität der Methoden (vgl. Döring und Bortz 2016, S. 502 f.).

Die gängigsten Methoden, die sich trotz der genannten Einschränkungen für ein Evaluationsvorhaben in der Wissenschaftskommunikation eignen, werden im Folgenden in vier Indikatorengruppen behandelt: Erregungsindikatoren, Emotionsindikatoren, Aufmerksamkeitsindikatoren und Bewertungsindikatoren.[1]

2 Erregungsindikatoren

Interessiert man sich dafür, welche Szenen eines Erklärfilms die Zielgruppe besonders anregen oder was beim Abendvortrag besonders aktivierend auf die Zuschauer:innen wirkt, kann man diese unspezifischen Reaktionen auf einen Reiz (Erregung) statt durch eine Befragung auch durch eine physiologische Messung

[1] Physiologische Messungen, die nur mittels medizinischer Großgeräte wie etwa einem MRT möglich sind (z. B. Untersuchungen des Hinstoffwechsels, vgl. Fahr und Hofer 2013, S. 358), werden hier explizit nicht besprochen.

in Erfahrung bringen. Die „sicherlich am häufigsten eingesetzte[n]“ (Fahr und Hofer 2013, S. 348) physiologischen Methoden im Kontext medienpsychologischer und kommunikationswissenschaftlicher Studien beziehen sich auf die elektrodermale Aktivität, d. h. „die Änderung der bioelektrischen Eigenschaften der Haut“ (Fahr 2013, S. 602). Diese basieren auf der Veränderung der Aktivität der sogenannten ekkrinen Schweißdrüsen, die vom sympathischen Teil des vegetativen Nervensystems gesteuert werden (vgl. Dawson et al. 2000, S. 203), welcher wiederum für die „nach außen gerichtete Handlungsbereitschaft“ (Fahr und Hofer 2013, S. 349) eines Menschen verantwortlich ist. Mit Veränderungen der elektrodermalen Aktivität korrelieren daher verschiedene psychologische Phänomene wie Aktivation[2], emotionale Reaktionen bzw. affektive Intensität, Aufmerksamkeit oder Informationsverarbeitung (vgl. Fahr 2013, S. 602; Fahr und Hofer 2013, S. 348).

Es lassen sich verschiedene Maße in diesem Zusammenhang unterscheiden[3], wie z. B. Hautpotenzial und Hautfeuchte, wobei die Untersuchung der Hautleitfähigkeit am etabliertesten ist (vgl. Döring und Bortz 2016, S. 518). Zur Messung werden Elektroden an der Innenseite der Hand befestigt[4], da sich dort – und an den Füßen – die höchste Dichte an ekkrinen Schweißdrüsen befinden (vgl. Birbaumer und Schmidt 2006). Eine Fremdspannung wird angelegt und der Hautwiderstand im Zeitverlauf aufgezeichnet. „Die typische Hautleitfähigkeitskurve besteht aus einem Grundniveau, das sich im Laufe der Medienrezeption erhöhen und verringern kann […]. Diese Grundlinie wird moduliert durch SCRs [Hautleitfähigkeitsreaktionen]“ (Fahr und Hofer 2013, S. 350). Diese Reaktionen sind reizbezogen und werden durch die Höhe des Ausschlags (Amplitude) sowie durch die Dauer und die Latenzzeit charakterisiert (vgl. Döring und Bortz 2016, S. 519).

Die Verarbeitung solcher Daten setzt ein hohes Maß an Fachkenntnis voraus. Nicht nur die Menge an potenziellen psychologischen Korrelaten, sondern auch die geringe Spezifität erschwert die eindeutige Interpretation. So lässt sich z. B. emotionale Erregung an den Messkurven der Versuchspersonen ablesen, nicht aber, ob diese Erregung positiv oder negativ ist (vgl. Fahr und Hofer 2013, S. 351).

[2] Mit Aktivation wird „eine sehr weit gefasste, ziemlich unspezifische Art der Erregung“ (Felser 2015, S. 88) bezeichnet.

[3] Für eine ausführliche Beschreibung verschiedener Hautphänomene vgl. Boucsein 2012.

[4] Für ein Schema der typischen Elektrodenplatzierungen vgl. Dawson et al. 2000, S. 205.

Zudem reagiert die Haut vergleichsweise langsam, wodurch Veränderungen der Hautleitfähigkeit zu einem schlechten Indikator werden, sobald die Stimuli zu schnell aufeinander folgen, da die „Zuordnung zwischen ‘Stimulus’ und Reaktion hier nicht mehr sinnvoll möglich“ (Fahr 2013, S. 603) ist. Dies betrifft viele multimodale Inhalte wie z. B. schnell geschnittene Informationsvideos.

Weitere klassische Erregungsindikatoren liefert die Untersuchung der kardiovaskulären Aktivität wie z. B. der Herzfrequenz, des Blutdrucks oder der peripheren Durchblutung. Sowohl Konzeption, Durchführung als auch die Auswertung solcher Messungen sind mit mindestens ähnlich großen Herausforderungen und Einschränkungen verbunden wie die der Hautleitfähigkeit. Details zu diesen Messungen finden sich zum Beispiel bei Fahr und Hofer (2013).

3 Emotionsindikatoren

Die Rezeption medialer Artefakte – vom Zeitungsartikel bis zum YouTube-Vortrag – hat immer auch eine emotionale Komponente, auch wenn deren Relevanz und Dominanz sehr unterschiedlich ausgeprägt sein kann. Dies gilt entsprechend auch für wissenschaftskommunikative Produkte und Maßnahmen, selbst dann, wenn aus Sicht der Produzent:innen der Schwerpunkt auf kognitiven Aspekten liegen mag – beispielsweise bei einer Ringvorlesung oder einem Erklärvideo zur Grundlagenforschung in der Physik. Emotionen – bzw. genauer emotionales Erleben – kann im Rahmen evaluatorischer Untersuchungen analysiert werden, dies erfordert jedoch in aller Regel detailliertes Fachwissen aufseiten der Evaluatoren und zudem vergleichsweise teures Spezialequipment. Insofern werden Emotionsindikatoren in der evaluatorischen Praxis nur in Ausnahmefällen bzw. im Rahmen größer angelegter Begleitforschungen herangezogen werden können.

Wenn emotionales Erleben mittels physiologischer Messungen erhoben werden soll, sind zwei Aktivitätsbereiche dabei von besonderem Interesse: Aktivitäten der Gesichtsmuskulatur und Aktivitäten der Augen, genauer noch des Lidschlags und der Pupille (vgl. Meinold 2005; Fahr 2013).

Zur Messung von Aktivitäten der Gesichtsmuskulatur wird klassisch – neben standardisierten Beobachtungen – die sog. Elektromyografie (EMG) eingesetzt. Dabei wird die elektrische Aktivität von Muskeln mittels Elektronen erfasst (vgl. Döring und Bortz 2016, S. 520 f.).[5] Es ist jedoch inzwischen auch möglich, zur

[5] Das entsprechende Verfahren zur Messung von Lidschlägen und Augenbewegungen wird als Elektrookulographie bezeichnet (vgl. Fahr 2013, S. 616).

Emotionserkennung mittels der Gesichtsmuskulatur spezifische Software einzusetzen, die auf Videos der Gesichter der Proband:innen angewendet wird. In beiden Fällen geht es um die Erfassung von – unwillkürlichen – Reaktionsmustern spezifischer Muskelgruppen, „die vergleichsweise valide und reliabel die Identifikation etwa diskreter Emotionen erlauben" (Fahr 2013, S. 614). Bereits in den 1980er Jahren wurde in der Psychologie ein Klassifikationsmodell entwickelt, das sechs Basisemotionen unterscheidet, die mithilfe der beschriebenen Verfahren detektiert werden können: glücklich, traurig, ärgerlich, überrascht, ängstlich und angewidert (vgl. Ekman und Friesen 1978). Wird hierbei nicht auf einschlägige Software zurückgegriffen, so handelt es sich bei einer solchen Kodierung um ein vergleichsweise aufwendiges Verfahren.

Bei der Beobachtung des menschlichen Auges ist in puncto Emotionsindikatorik neben dem Lidschlag (vgl. Meinold 2005) die Pupillengröße von besonderem Interesse. Diese kann nicht intendiert verändert werden und wird nicht nur von der Lichtmenge, die auf sie einwirkt, sondern auch von psychischen Faktoren wie beispielsweise mentaler Belastung oder affektivem Interesse beeinflusst (vgl. Fahr 2013, S. 617). Insbesondere gilt: „Bei besonderer Aktivität des Limbischen Systems (Angst, Erregung oder hohe mentale Belastung) erweitert sich die Pupille, bei psychischer Überforderung verkleinert sie sich" (Fahr 2013, S. 617). Zur Messung der Pupillengröße werden videobasierte Eyetracker eingesetzt (vgl. Abschnitt Aufmerksamkeitsindikatoren). Genauso wie beim Indikator Gesichtsmuskulatur sind zur Messung von Lidschlagfrequenz und Pupillengröße spezifische, kostenintensive Geräte und geschulte Mitarbeiter:innen zur Durchführung und Auswertung der Messungen notwendig.

4 Aufmerksamkeitsindikatoren

Wenn es um die Messung der Aufmerksamkeit von Menschen beim Umgang mit medialen oder kommunikativen Artefakten geht, spielen Blickbewegungen eine zentrale Rolle. Die Methode der Blickaufzeichnung wird in der empirischen Medienforschung inzwischen vielfach erfolgreich eingesetzt (vgl. Bucher und Schumacher 2006; Gehl 2013; Geise 2011b; Holsanova et al. 2009; Niemann und Krieg 2011; Schumacher 2009). In der Wissenschaftskommunikationsforschung sind Studien, die dieses Verfahren einsetzen, bisher selten (vgl. Kessler 2021), aber durchaus zu finden (vgl. z. B. Böhmert et al. 2021; Bucher und Niemann 2012, 2015; Kessler und Zillich 2019). Mit der Methode lässt sich strategisches, intentionales, aber auch nicht-intentionales Verhalten im Rezeptionsprozess offenlegen (vgl. Schumacher 2009, S. 110).

Die Methode der Blickaufzeichnung[6] kommt unmittelbar während des Rezeptionsprozesses von Versuchspersonen zum Einsatz. Im Unterschied zu allen Formen des self-reporting ermöglicht sie „einen direkten Einblick in die Interaktion zwischen Stimulus und Rezipient" (Bucher et al. 2010, S. 385) und hat dabei deutlicher weniger mit den bekannten Problemen der Reaktivität und Validität zu kämpfen: Blickdaten gelten als „weitestgehend authentisch" (Bucher 2011, S. 116). Grund dafür ist, dass „Blickbewegungen zu einem hohen Maße als nicht-intentional bezeichnet werden können" (Gehl 2013, S. 163)[7].

Blickbewegungen sind Indikatoren für die Allokation von Aufmerksamkeit seitens Proband:innen (vgl. Bente 2004, S. 298)[8]. Sie sind zum Verständnis von Rezeptionsprozessen in vielfacher Hinsicht hilfreich, geben sie doch Auskunft über:

- die Selektionsstrategien der Proband:innen (was wird überhaupt angesehen?),
- den Grad von Aufmerksamkeit und Interesse hinsichtlich einzelner Elemente des Stimulus (Dauer der Betrachtung),
- Rezeptionssequenzen und damit gleichzeitig über Erschließungsstrategien der Proband:innen und
- die Qualität der Rezeption (z. B. Scannen vs. Lesen)
  (vgl. Bucher et al. 2010, S. 385).

Damit erlauben sie konkrete „Rückschlüsse auf […] Aneignungshandlungen" (Gehl 2013, S. 164) von Proband:innen. Der Auge-Geist-Hypothese (Just und Carpenter 1980) und der kritischen Reflexion dieses Ansatzes (Geise 2011a; Schumacher 2012) folgend, lassen darüber hinaus die mit einem Eyetracker gemessenen Fixationen von Elementen – Buchstaben, Bilder etc. – Rückschlüsse auf eine mögliche kognitive Verarbeitung zu.

[6] Die Darstellung in diesem und dem folgenden Absatz ist angelehnt an bzw. übernommen aus Niemann 2015, S. 73–75.

[7] Völlig frei von Reaktivitätseffeken ist die Methode dennoch nicht. Je nach Anwendungsfall können sich diese schon allein aus der Tatsache ergeben, dass Proband:innen eine spezielle Brille tragen müssen bzw. ihre Sitzposition vor dem Stimuluscomputer während der Rezeptionsphase nicht uneingeschränkt verändern können.

[8] Schumacher weist in diesem Zusammenhang richtig auf die Problematiken des extrafovealen Sehens und der *covert attention* hin, die dafür verantwortlich sind, dass nicht in jedem Fall von einer völligen Übereinstimmung von Blickbewegungen und der Verschiebung der kognitiven Aufmerksamkeit ausgegangen werden kann (vgl. Schumacher 2009, S. 110 f.).

Trotz der genannten Potenziale der Methode gilt der Satz: „Zu wissen, was ein Proband anschaut, heißt noch nicht zu wissen, was er sieht" (Schumacher 2012, S. 115). Blickdaten müssen kontextualisiert werden, um Aussagen zur tatsächlichen kognitiven Verarbeitung von Stimuli machen zu können (vgl. Bucher 2012, S. 265; Gehl 2013, S. 164 f.; Schumacher 2012, S. 115 f.). Wird darauf verzichtet, so „bleibt beispielsweise unklar, ob ein Seitenelement überdurchschnittlich lange fixiert wurde, weil es dem Betrachter besonders interessant erschien, ober ob an dieser Stelle Rezeptionsschwierigkeiten vorlagen" (Gehl 2013, S. 164). Nur durch Methodenkombination, etwa mit Lautem Denken[9] und der Aufzeichnung nonverbaler Handlungen sind „schlüssige Erklärungen für bestimmte Rezeptionsabfolgen" (Schumacher 2012, S. 115) ableitbar.

In der Forschungspraxis werden anwendungsbezogene Blickaufzeichnungen heute in aller Regel mit videobasierten Eyetrackern durchgeführt.[10] Dabei kommen – je nach Untersuchungsgegenstand – bevorzugt non-invasive, häufig unterhalb des Proband:innenmonitors installierte bzw. sogar in diesen integrierte Geräte zum Einsatz, die das Verhalten der Rezipient:innen während einer Rezeptionssituation so gut wie nicht beeinflussen. Geht es um Untersuchungssituationen, in denen die relevanten Inhalte nicht auf einem Computermonitor präsentiert werden können (Live-Veranstaltungen, Handyinhalte, Elemente im Straßenraum etc.) werden sog. Headmounted-Systeme verwendet. Diese arbeiten ebenfalls videobasiert, zeichnen jedoch zusätzlich über eine Frontkamera auf, was eine Versuchsperson sieht, damit die erhobenen Blickdaten mit der Situation vor dem Gesichtsfeld in Bezug gesetzt werden können. Inzwischen sind derartige Geräte sehr miniaturisiert verfügbar und ähneln einer handelsüblichen Brille mit breitem Rahmen.

In der Wissenschaftskommunikationsforschung werden Blickaufzeichnungen fast immer mit einer evaluatorischen Perspektive auf kommunikative Artefakte – seien es nun Live-Formate, Videos, Printprodukte oder auch eine VR-Anwendung – eingesetzt. Andererseits führen die vergleichsweise hohen technischen, zeitlichen und finanziellen Anforderungen an eine Untersuchung mittels Blickaufzeichnung sowie die erforderlichen Fachkenntnisse bei der Anwendung und

[9] Mit der Methode des Lauten Denkens kann individuelles Verstehen erhoben werden (vgl. Schumacher 2009, S. 105). Dazu werden Proband:innen gebeten, während eines Rezeptionsprozesses laut zu verbalisieren, was ihnen durch den Kopf geht. Zu Vor- und Nachteilen sowie Varianten der Methode vgl. Bilandzic 2017.

[10] Zu detaillierten (technischen) Grundlagen der Blickaufzeichnung vgl. Geise 2011a; Bente 2004.

Auswertung der Methode dazu, dass sie im Repertoire reiner Praxis- oder typischer Selbstevaluationen nicht zu finden ist. Praxisbezogene Fragestellungen, zu deren Beantwortung die Methode einen relevanten Beitrag leisten kann, sind beispielsweise:

- Fragen nach der Relevanz *wissenschaftlichen Wissens* für Rezipient:innen: Damit wissenschaftliches Wissen verstanden oder gar behalten werden kann, muss es zunächst von Rezipient:innen wahrgenommen werden. Blickaufzeichnungen können Auskunft darüber geben, inwiefern Passagen/Bereiche/Abschnitte in Texten, Präsentationen, Videos etc., in denen wissenschaftliches Wissen vermittelt werden soll, von Rezipient:innen überhaupt mit Aufmerksamkeit bedacht werden, und wenn ja, mit wie viel – auch und gerade im Verhältnis zum Rest eines medialen Produkts.
- Fragen nach der Relevanz von *Personen bzw. von spezifischen Personen* für Rezipient:innen: Wir alle wissen aus Erfahrung, dass es einen erheblichen Unterschied macht, wie genau wissenschaftliche Inhalte an uns herangetragen werden, und zwar sowohl für das Unterhaltungsempfinden als auch für die Bereitschaft, sich tatsächlich mit den Inhalten kognitiv auseinanderzusetzen. Ein wesentlicher Faktor ist dabei der Mensch: Benötige ich das Bild einer vortragenden Person bei einem Präsentationsvideo, oder ist das egal? Spielt es eine Rolle, ob Gesten und Bewegungen der Person auf dem Vortragsvideo neben den Folien noch erkennbar sind? Ändert sich die Aufmerksamkeit der Rezipient:innen für eine vortragende Person beim Tag der offenen Tür, wenn es sich dabei um eine Professorin handelt und nicht um eine Doktorandin, selbst wenn diese nicht lustiger, lauter oder performativer den identischen Inhalt vorträgt? Zu all diesen Fragen können Blickaufzeichnungen ein empirisches Fundament bereitstellen.
- Fragen nach der Relevanz von *Unterhaltung* für Rezipient:innen: Auch zur Klärung dieser Grundsatzfrage der Wissenschaftskommunikation können Blickaufzeichnungen etwas beitragen, indem durch den Einsatz der Methode beispielsweise deutlich wird, wie sich die Aufmerksamkeit von Rezipient:innen etwa bei einem Science-Slam auf die unterhaltenden und die wissenschaftlichen Elemente der Präsentationsfolien verteilt.
- Grundsätzliche Fragen der *Usability* in wissenschaftskommunikativen Produkten: Hierbei handelt es sich um kein Spezifikum der Wissenschaftskommunikation, aber für alle medialen Produkte gilt: Nur wenn diese ohne nennenswerte Kommunikationsprobleme genutzt werden können, besteht überhaupt die Möglichkeit, dass die von ihren Produzent:innen anvisierten kommunikativen Ziele erreicht werden. Dabei kann die Methode der Blickaufzeichnung hilfreich

sein, wird durch ihren Einsatz doch beispielsweise deutlich, wie Navigationspfade auf Webseiten verlaufen, welche Bereiche eines Flyers oder einer Broschüre gar nicht wahrgenommen werden oder welcher Button in einem Computerspiel wegen unklarer Beschriftung viel länger angesehen wird, als es funktional sinnvoll wäre.

5 Bewertungsindikatoren

Sei es das Unterhaltungsempfinden während eines Science-Slams oder die Einschätzung der Verständlichkeit von Expert:innen während einer Debatte: In beiden Fällen geht es um Bewertungen, die sich über die Zeit verändern können und deren genaue Kenntnis zu jedem Zeitpunkt Rückschlüsse auf einzelne Bestandteile z. B. eines Vortrags erlaubt.

Als eine Art „X-ray of the program as experienced by the audience“ beschreibt Millard (1992, S. 1) das Anfang der 1940er Jahre von den Forschern Lazarsfeld und Stanton entwickelte Verfahren „zur Evaluation von im Radio ausgestrahlten Unterhaltungsprogrammen“ (Maier 2013, S. 172), das er noch als Direct Measurement of Audience Response bezeichnet. Bis heute haben sich zwei andere Namen für die Methode durchgesetzt: Real-Time-Response (RTR)-Messung und Continuous-Response-Messung (vgl. Maurer 2013, S. 220). Im deutschsprachigen Raum scheint der Ausdruck RTR-Messung gebräuchlicher zu sein (vgl. Waldvogel und Metz 2017, S. 308) und wird daher auch in diesem Kapitel verwendet.

„RTR-Messungen ermöglichen die rezeptionsbegleitende Erfassung individueller und subjektiver Reaktionen auf audio-visuelle Stimuli“ (Waldvogel und Metz 2017, S. 308). Ihre Eindrücke können Versuchspersonen dabei über eigens dafür entwickelte physische oder (heute üblicher) virtualisierte (d. h. Smartphone oder Computer) Eingabegeräte zu jedem Zeitpunkt der Rezeption kontinuierlich bzw. quasi-kontinuierlich[11] zurückmelden. Die übermittelte Selbstauskunft kann dabei eine Vielzahl von Konstrukten abdecken, wie z. B., ob ein Video als angenehm empfunden wird, eine Vorlesung als informativ oder welcher der beiden politischen Kandidaten in einem TV-Duell gerade den besseren Eindruck macht (vgl. Waldvogel und Metz 2017, S. 308).

[11] Nur wenige physische Eingabegeräte könnten technisch überhaupt in der Lage sein, tatsächlich kontinuierliche Daten zu erzeugen, und virtualisierte Systeme besitzen generell keinen stetigen Datenstrom, sondern sind immer diskret.

Tab. 1 Übersicht über die verschiedenen Typen von RTR-Eingabesystemen

Gerätetyp	Knopf (Push-Button)	Drehregler (Dialer)	Schieberegler (Slider)	Spielehebel (Joystick)
Datenniveau	Bei einem Knopf: nominal (dichotom). Bei mehreren Knöpfen auch nominal bis quasi-metrisch möglich.	Ordinal/quasi-metrisch	Ordinal/quasi-metrisch	Quasi-metrisch
Anzahl Bewertungs-items	1	1	1	1–2

Tab. 1 gibt eine Übersicht über die verschiedenen Typen von RTR-Eingabesystemen[12]:

RTR-Messungen bieten einige Vorteile gegenüber postrezeptiven Verfahren wie Befragungen oder anderen rezeptionsbegleitenden Verfahren wie Lautem Denken. Letzteres bedeutet für gewöhnlich einen erheblichen Eingriff in die Rezeptionssituation, da die laut geäußerten Eindrücke der Versuchspersonen die Rezeption entweder unterbrechen oder überlagern. Im Vergleich hierzu können Rückmeldungen per RTR-Eingabegerät schnell und nebenbei abgegeben werden. Gegenüber den postrezeptiven Verfahren haben RTR-Messungen den Vorteil, dass sich Versuchspersonen nicht erst an Rezeptionssituationen erinnern müssen, ihre Beurteilungen nicht nachträglich rationalisiert werden können und weniger starke Verzerrungen aufgrund von sozialer Erwünschtheit auftreten (vgl. Taddicken et al. 2020, S. 56). Hinzu kommt, dass die „Erfassung von Eindrücken einzelner (medialer) Inhalte mittels RTR-Messung […] dezidiert auf konkrete Merkmale des gezeigten Materials“ (Taddicken et al. 2020, S. 56) zurückgeführt werden kann. Beispielsweise könnte man mittels einer RTR-Messung herausfinden, welche Aspekte eines Präsentationsfilms einer Hochschule zu einem naturwissenschaftlichen Thema sich positiv oder negativ auf die von den Zuschauer:innen zugeschriebene Kompetenz der oder des Präsentierenden auswirkt.

[12] Es handelt sich um eine angepasste Darstellung der Tabelle bei Waldvogel und Metz (2017, S. 310). Für einen visuellen Eindruck verschiedener physischer Eingabegeräte vgl. Ottler 2013, S. 117.

Probleme von RTR-Messungen sind die mitunter stark variierenden interindividuellen Schwellwerte, ab wann ein Knopf zu betätigen bzw. ein Regler zu verschieben ist (vgl. Waldvogel und Metz 2017, S. 311). Außerdem liegt die Vermutung nahe, dass „die ungewohnte Selbstauskunft die Introspektion erhöhen könnte, was die Übertragbarkeit auf alltägliche, unreflektierte Situationen erschweren kann" (Waldvogel und Metz 2017, S. 314). Damit einhergehend verlangen RTR-Messungen bei den Versuchspersonen zusätzliche kognitive Ressourcen während der Rezeption – wie bei allen rezeptionsbegleitenden Verfahren. Stehen diese nicht zur Verfügung, weil die Versuchsperson „etwa von einem Inhalt besonders gebannt [ist], ist es denkbar, dass ‚vergessen' wird, dem Erleben Ausdruck zu verleihen" (Fahr 2013, S. 620).

Auch wenn finanzielle Vorbehalte gegen RTR-Verfahren mit physischen Messgeräten dank der zunehmenden Virtualisierung des Verfahrens (Waldvogel und Metz 2017, S. 316) fast obsolet geworden sind, bleibt es dabei, dass zur Erhebung und Auswertung spezielle Software benötigt wird.

Für Evaluationsvorhaben ist in erster Linie die stark begrenzte Anzahl an Items, über die eine RTR-Messungen Auskunft geben kann, ein Nachteil. Zum Beispiel kann eine Versuchsperson nicht zwei Schieberegler bedienen und gleichzeitig Auskunft über Unterhaltsamkeit und Verständlichkeit eines Vortrags geben. Außerdem muss man bedenken, dass RTR-Messungen zunächst nichts darüber aussagen, weshalb Versuchspersonen bestimmte Bewertungen vorgenommen haben. Begründungen der abgegebenen Bewertungen lassen sich nur durch Kombinationen mit anderen Methoden wie zum Beispiel Befragungen oder Lautem Denken erheben.

6 Fazit

Im Vergleich zu anderen Evaluationsmethoden wird deutlich, dass der Einsatz physiologischer Messungen sowohl in Konzeption, Durchführung und Auswertung komplex ist und nicht ohne Erfahrung auskommt. Zudem ist beinahe immer eine Methodenkombination erforderlich, um Begründungszusammenhänge zu erfassen.

Dennoch haben die aufgeführten Methoden ihre jeweiligen spezifischen Vorteile und können ein sehr tiefgehendes Verständnis für Rezeptionsprozesse bei wissenschaftskommunikativen Artefakten schaffen, die eine Evaluation erheblich bereichern können.

Genau deswegen sollte man sich nicht davon abhalten lassen, auch solche elaborierteren Methoden für eine Evaluation in Erwägung zu ziehen, auch wenn dies in der Praxis häufig bedeutet, dass externe Expertise hinzugezogen werden muss – etwa durch die Kooperation mit einer Forschungseinrichtung.

Literatur

Bente G (2004) Erfassung und Analyse des Blickverhaltens. In: Bente G, Mangold R, Vorderer P (Hrsg) Lehrbuch der Medienpsychologie. Hogrefe, Göttingen, S 297–324

Birbaumer N, Schmidt RF (2006) Biologische Psychologie, 7. Aufl. Springer, Heidelberg. https://doi.org/10.1007/978-3-540-95938-0

Bilandzic H (2017) Lautes Denken. In: Mikos L, Wegener C (Hrsg) Qualitative Medienforschung: Ein Handbuch, 2. Aufl. UVK, Konstanz, S 406–413

Böhmert C, Niemann P, Hansen-Schirra S, Nitzke J (2021) Wen verstehen wir besser? Eine vergleichende Rezeptionsstudie zu Kurzmeldungen von Journalisten und Wissenschaftlern. In: Milde J, Vogel IC, Dern M (Hrsg) Intention und Rezeption von Wissenschaftskommunikation, Herbert von Halem Verlag, Köln, S 110–126

Boucsein W (2012) Electrodermal activity. 2. Aufl. Springer, Heidelberg. https://doi.org/10.1007/978-1-4614-1126-0

Bucher HJ (2011) »Man sieht, was man hört« oder: Multimodales Verstehen als interaktionale Aneignung. Blickaufzeichnungsstudie zur audiovisuellen Rezeption. In: Schneider JG, Stöckl H (Hrsg) Medientheorien und Multimodalität. Ein TV-Werbespot - Sieben methodische Beschreibungsansätze. Herbert von Halem Verlag, Köln, S 109–150

Bucher HJ (2012) Intermodale Effekte in der audio-visuellen Kommunikation: Blickaufzeichnungsstudie zur Rezeption von zwei Werbespots. In: Bucher H-J, Schumacher P (Hrsg) Interaktionale Rezeptionsforschung: Theorie und Methode der Blickaufzeichnung in der Medienforschung. Springer VS, Wiesbaden, S 257–296. https://doi.org/10.1007/978-3-531-93166-1_10

Bucher HJ, Niemann P (2012) Visualizing science: the reception of powerpoint presentations. Vis Commun 11(3):283–306. https://doi.org/10.1177/1470357212446409

Bucher HJ, Niemann P (2015) Medialisierung der Wissenschaftskommunikation: Vom Vortrag zur multimodalen Präsentation. In: Schäfer MS, Kristiansen S, Bonfadelli H (Hrsg) Wissenschaftskommunikation im Wandel. Herbert von Halem Verlag, Köln, S 68–101

Bucher HJ, Schumacher P (2006) The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media. Communications 31 (3):347–368. https://doi.org/10.1515/COMMUN.2006.022

Bucher HJ, Krieg M, Niemann P (2010) Die wissenschaftliche Präsentation als multimodale Kommunikationsform: zur Rezeption von Powerpoint- Vorträgen. In: Bucher HJ, Gloning T, Lehnen K (Hrsg) Neue Medien – neue Formate: Ausdifferenzierung und Konvergenz in der Medienkommunikation. Campus, Frankfurt a. M. [u. a.], S 375–406

Dawson ME, Schell AM, Filion DL (2000) The electrodermal system. In: Cacioppo JT, Tassinary LG, Berntson GG (Hrsg) Handbook of Psychophysiology, 2. Aufl. Cambridge University Press, Cambridge

Döring N, Bortz J (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-41089-5
Ekman P, Friesen WV (1978) Facial action coding system. Consulting Psychologists Press, Palo Alto, A technique for the measurement of facial movement
Fahr A (2013) Physiologische Ansätze der Wirkungsmessung. In: Schweiger W, Fahr A (Hrsg) Handbuch Medienwirkungsforschung, Springer, Wiesbaden. https://doi.org/10.1007/978-3-531-18967-3_32
Fahr A, Hofer M (2013) Psychophysiologische Messmethoden. In: Möhring W, Schlütz D (Hrsg) Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft. Springer, Wiesbaden, S 347–365. https://doi.org/10.1007/978-3-531-18776-1_19
Felser G (2015) Werbe- und Konsumentenpsychologie. 4. Aufl. Springer, Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37645-0
Gehl D (2013) Vom Betrachten zum Verstehen. Die Diagnose von Rezeptionsprozessen und Wissensveränderungen bei multimodalen Printclustern. Springer, Wiesbaden. https://doi.org/10.1007/978-3-531-19823-1
Geise S (2011a) Eyetracking in der Kommunikations- und Medienwissenschaft: Theorie, Methode und kritische Reflexion. Studies in Communication and Media 1(2):149–263. https://doi.org/10.5771/2192-4007-2011-2-149
Geise S (2011b) Vision that matters. VS Verlag, Wiesbaden, Die Funktions- und Wirkungslogik visueller politischer Kommunikation am Beispiel des Wahlplakats. https://doi.org/10.1007/978-3-531-92736-7
Holsanova J, Holmberg N, Holmqvist K (2009) Reading information graphics: The role of spatial contiguity and dual attentional guidance. Appl Cogn Psychol 23(9):1215–1226. https://doi.org/10.1002/acp.1525
Just MA, Carpenter PA (1980) A theory of reading: From eye fixations to comprehension. Psychol Rev 87(4):329–354. https://doi.org/10.1037/0033-295X.87.4.329
Kessler SH (2021) The use of the eye tracking method in science communication research. Paper accepted for the Dreiländertagung für Kommunikationswissenschaft, Zürich (virtual conference), 7.-9. April
Kessler SH, Zillich AF (2019) Searching online for information about vaccination: Assessing the influence of user-specific cognitive factors using eye-tracking. Health Communication 34(10):1150–1158. https://doi.org/10.1080/10410236.2018.1465793
Millard WJ (1992) A History of Handsets for Direct Measurement of Audience Response. In: International Journal of Public Opinion Research 4(1):1–17. https://doi.org/10.1093/ijpor/4.1.1
Maier J (2013) Rezeptionsbegleitende Erfassung individueller Reaktionen auf Medieninhalte. Bedeutung, Varianten, Qualität und Analyse von Real-Time-Response-Messungen. ESSACHESS – Journal for Communication Studies 6(1):169–184
Maurer M (2013) Real-Time Response Messung: Kontinuierliche Befragung in Echtzeit. In: Möhring W, Schlütz D (Hrsg) Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft. Springer, Wiesbaden, S 219–234. https://doi.org/10.1007/978-3-531-18776-1_12
Meinold PE (2005) Psychologie des Lidschlags - eine literatur- und methodenkritische Studie, Inaugural-Dissertation an der Philosophischen Fakultät der Universität zu Köln

Niemann P (2015) Die Pseudo-Medialisierung des Wahlkampfs: Eine rezipientenorientierte Analyse zweier Onlinewahlkämpfe politischer Parteien. Springer, Wiesbaden. https://doi.org/10.1007/978-3-658-07610-8

Niemann P, Krieg M (2011) Von der Bleiwüste bis zur Diashow: Zur Rezeption zentraler Formen wissenschaftlicher Präsentationen. Zeitschrift für angewandte Linguistik 54(1):111–143. https://doi.org/10.1515/zfal.2011.006

Ottler S (2013) RTR-Messung: Möglichkeiten und Grenzen einer sozialwissenschaftlichen Methode. In: Bachl M, Brettschneider F, Ottler S (Hrsg) Das TV-Duell in Baden-Württemberg 2011. Inhalte, Wahrnehmungen und Wirkungen. Springer, Wiesbaden, S 113–134. https://doi.org/10.1007/978-3-658-00792-8_6

Schumacher P (2009) Rezeption als Interaktion: Wahrnehmung und Nutzung multimodaler Darstellungsformen im Online-Journalismus. Nomos, Baden-Baden. https://doi.org/10.5771/9783845216911

Schumacher P (2012) Blickaufzeichnung in der Rezeptionsforschung: Befunde, Probleme und Perspektiven. In: Bucher HJ, Schumacher P (Hrsg) Interaktionale Rezeptionsforschung: Theorie und Methode der Blickaufzeichnung in der Medienforschung. Springer, Wiesbaden, S 111–134. https://doi.org/10.1007/978-3-531-93166-1_4

Taddicken M, Wicke N, Willems K (2020) Verständlich und kompetent? Eine Echtzeitanalyse der Wahrnehmung und Beurteilung von Expert*innen in der Wissenschaftskommunikation. M&K Medien & Kommunikationswissenschaft 68(1–2):50–72. https://doi.org/10.5771/1615-634x-2020-1-2-50

Waldvogel T, Metz T (2017) Real-Time-Response-Messungen. In: Jäckle S (Hrsg) Neue Trends in den Sozialwissenschaften. Innovative Techniken für qualitative und quantitative Forschung. Springer, Wiesbaden, S 307–331. https://doi.org/10.1007/978-3-658-17189-6_11

Philipp Niemann ist stellvertretender Direktor und wissenschaftlicher Leiter des Nationalen Instituts für Wissenschaftskommunikation (NaWik). Zuvor war er als Nachwuchsgruppenleiter im Department für Wissenschaftskommunikation am Karlsruher Institut für Technologie (KIT) tätig. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftskommunikation, qualitative Rezeptionsforschung und politische Kommunikation.

Yannic Scheuermann ist wissenschaftlicher Mitarbeiter am Nationalen Institut für Wissenschaftskommunikation (NaWik). Dort beschäftigt er sich mit der Evaluation wissenschaftskommunikativer Aktivitäten. Seine Forschungsinteressen liegen im Bereich qualitativer Methoden und der Entwicklung technischer Lösungen für kommunikationswissenschaftliche Fragestellungen.


cc
BY



Praxisbeitrag: Physiologische Messungen in der evaluatorischen Praxis

Eine Eyetracking-Studie zum virtuellen KATRIN-Experiment

Christian Humm und Philipp Niemann

Zusammenfassung

In diesem Beitrag wird anhand der Untersuchung einer Virtual Reality Umgebung aus dem Themenbereich der Physik – KATRIN VR (vr.nawik.de) – beispielhaft gezeigt, wie eine spezifische Methode aus dem Repertoire der Verfahren zur physiologischen Messung bei der Evaluation von Angeboten der Wissenschaftskommunikation eingesetzt werden kann: Die Methode der Blickaufzeichnung. Dazu werden in einem ersten Schritt relevante evaluationsbezogene Fragestellungen am Beispiel von KATRIN VR aufgezeigt, die anschließend in einer quasi-experimentellen multimethodischen Untersuchung adressiert werden. Dabei wird die im Vordergrund stehende Blickaufzeichnung mit einem Wissenstest und der Methode des Lauten Denkens kombiniert, um Rezeption und Wirkungsaspekte der Virtual Reality Umgebung zu erfassen.

C. Humm (✉)
Büro des Universitätspräsidenten, Universität des Saarlandes, Saarbrücken, Saarland, Deutschland
E-Mail: christian.humm@uni-saarland.de

P. Niemann
Nationales Institut für Wissenschaftskommunikation gGmbH, Heidelberg, Deutschland
E-Mail: niemann@nawik.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_13

1 Einleitung

In diesem Beitrag soll anhand der Untersuchung einer Virtual Reality Umgebung aus dem Themenbereich der Physik – KATRIN VR (vr.nawik.de) – beispielhaft gezeigt werden, wie eine spezifische Methode aus dem Repertoire der Verfahren zur physiologischen Messung bei der Evaluation von Angeboten der Wissenschaftskommunikation eingesetzt werden kann: Die Methode der Blickaufzeichnung.

Blickdaten werden im Folgenden primär als Aufmerksamkeitsindikatoren herangezogen (siehe auch Niemann und Scheuermann in diesem Band). Ziel des Beitrags ist es, am konkreten Untersuchungsgegenstand einer VR-Umgebung (Abschn. 2) zunächst evaluative Forschungsfragen aufzuzeigen und zu diskutieren, die mithilfe von Blickdaten beantwortet werden können (Abschn. 3). Daran anschließend fokussiert sich der Beitrag auf die Aspekte, die in einer Studie des Forschungsprojekts "Science In Presentations" tatsächlich untersucht wurden, zeigt das methodische Gesamtsetting dieser Untersuchung auf (Abschn. 4) und präsentiert schließlich ausgewählte Auswertungen der erhobenen Daten (Abschn. 5). Neben der Zusammenfassung der Ergebnisse werden in Abschn. 6 auch die Limitationen der Methode der Blickaufzeichnung im konkreten Anwendungsfall sowie die forschungspraktischen Herausforderungen beim Einsatz des Verfahrens noch einmal expliziert. Damit kann eine sinnvolle Entscheidungsgrundlage für den Einsatz von Blickaufzeichnungen in der evaluatorischen Praxis bereitgestellt werden.

2 Die VR-Umgebung zum KATRIN-Experiment

Das KATRIN-Experiment ist ein groß angelegtes physikalisches Experiment mit dem Ziel, die absolute Masse von Neutrinos zu bestimmen. Die gesamte Versuchsanlage ist 70 m lang und befindet sich am Campus Nord des Karlsruher Instituts für Technologie (KIT) in der Nähe von Karlsruhe.

Als Teil des Forschungsprojekts "Science In Presentations"[1] wurde eine 360°- und VR-Umgebung des Experiments erstellt. In dieser multimodalen Umgebung können die Nutzer:innen das KATRIN-Experiment und seine Gerätschaften auf

[1] In dem Projekt wurden von 2016 bis 2021 verschiedene Präsentationsformen der Wissenschaftskommunikation untersucht, weitere Informationen unter www.science-in-presentations.de.

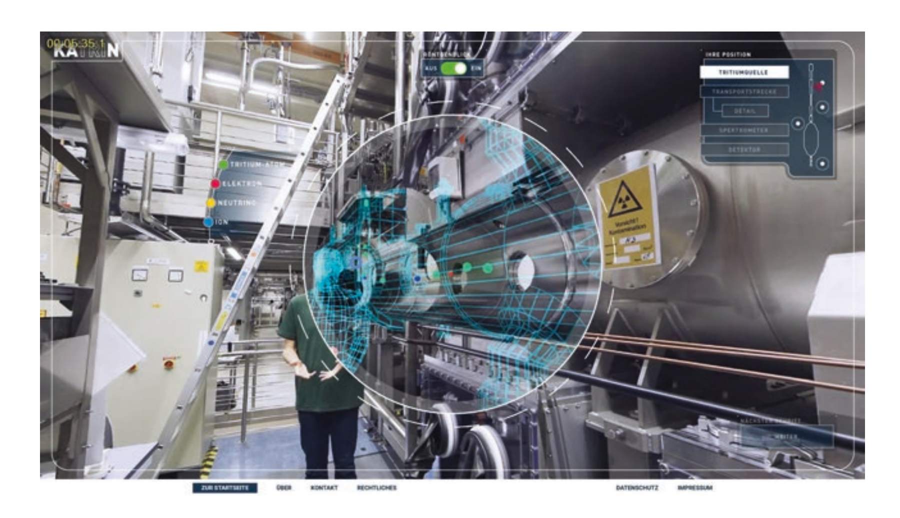


Abb. 1 Standfoto der Tour mit aktiviertem Röntgenblick. (Quelle: KATRIN VR)

zwei verschiedene Arten erkunden: Zum einen können sie die Umgebung frei und ohne Führung explorieren. Zum anderen können die Nutzer:innen an einer geführten Tour teilnehmen, die sie in die Kernkonzepte des Experiments einführt. Die Tour dauert etwa 15 min. Der Guide ist ein Physiker, der zum KATRIN-Experiment promoviert hat.

Bei mehreren Gelegenheiten während dieser Führung kann man eine sogenannte Röntgenansicht einschalten und mit Gerätschaften aus dem Experiment interagieren (vgl. Abb. 1). So ist es beispielsweise möglich, Spannungen zu manipulieren und in einer Animation zu sehen, wie dies die Elementarteilchen im Experiment beeinflusst. Während der Tour werden zudem zwei animierte Erklärfilme abgespielt.

3 Relevante evaluationsbezogene Forschungsfragen

Die vorgestellte VR-Umgebung eignet sich aufgrund ihrer modalen Komplexität gut, um eine Vielzahl von evaluativen Fragestellungen zu untersuchen, bei denen die Methode der Blickaufzeichnung grundsätzlich hilfreich sein kann (siehe auch Niemann und Scheuermann in diesem Band). Auswertungen der *Selektions-*

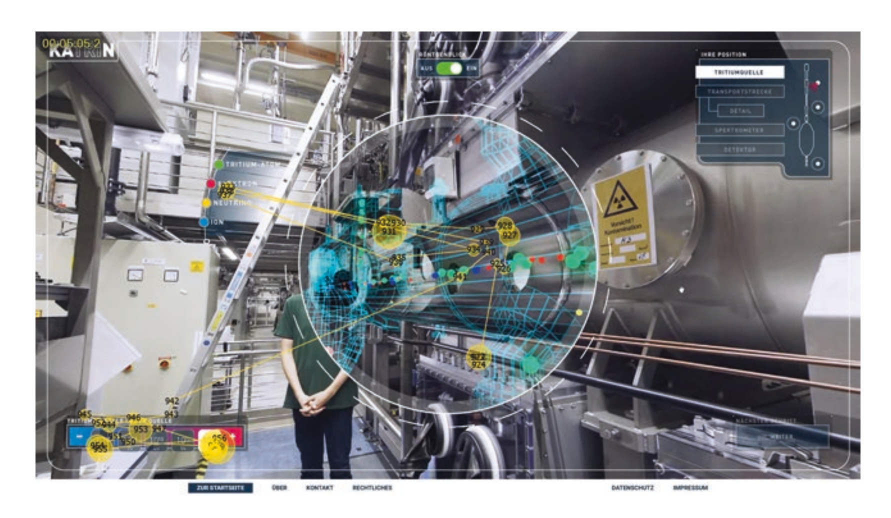


Abb. 2 Scanpath[2] eines Teilnehmers (hellblau, Dauer: 10 s). Je größer der Kreis, desto länger die Fixation der Augen. Die angegebenen Zahlen zeigen die Reihenfolge der Betrachtung. (Quelle: eigene Darstellung)

strategien können hier sowohl hinsichtlich der *Relevanz wissenschaftlichen Wissens* als auch der *Relevanz von spezifischen Personen* Auskunft geben: Werden zum Beispiel die einzelnen Elementarteilchen, von denen die Rede ist, überhaupt angesehen, oder befassen sich Nutzer:innen eher mit der großen Forschungsanlage oder anderen Elementen im Raum? Wie oft und wann erhält der Physiker im Raum visuelle Aufmerksamkeit? Die Untersuchung des *Grades von Aufmerksamkeit und Interesse* bietet sich ebenfalls besonders mit Blick auf die fachwissenschaftlichen Elemente der VR-Umgebung an: Wie lange werden diese im Vergleich zu anderen Elementen betrachtet (vgl. Abb. 2)?

Bei neuartigen medialen Artefakten wie einer VR-Umgebung spielt immer auch die Frage nach der generellen Nutzbarkeit und der spezifischen *Usability* eine Rolle. In diesem Zusammenhang können die mittels Blickaufzeichnungen sichtbar werdenden *Rezeptionssequenzen* von Nutzer:innen herangezogen werden, geben sie doch zugleich Auskunft über angewandte *Erschließungsstrategien*: In welcher Reihenfolge werden die verschiedenen

[2] Scanpath bezeichnet die visuelle Darstellung der Blickabfolge während einer bestimmten Zeitdauer. Einen detaillierten Einblick in Analyse- und Auswerteverfahren von Blickaufzeichnungen im Medienkontext bietet Geise (2011).

Elemente der VR-Umgebung betrachtet? Werden etwa im Anschluss an die visuelle Erschließung von Tasten auf dem virtuellen Tablet diese Tasten auch entsprechend ihrer intendierten Funktion genutzt (vgl. ebenfalls Abb. 2)?

Auch die *Qualität bzw. die Art der Rezeption* spielt im vorgestellten Beispiel eine Rolle, etwa wenn es darum geht, ob der Text in den virtuellen Infokästen von Nutzer:innen tatsächlich gelesen wird. Hier können Blickdaten unmittelbar Auskunft geben, da sich das Blickmuster, das beim Lesen eines Textes entsteht, deutlich von dem unterscheidet, das man beim Überfliegen des Textes erhält.

In der nachfolgend im Detail vorgestellten Studie werden einige der hier skizzierten Fragestellung mit der Methode der Blickaufzeichnung adressiert.

4 Methodisches Vorgehen in der Studie

Virtual Reality und 360°-Anwendungen sind relative neue Werkzeuge für die (Hochschul-) Bildung. In einem Literaturreview kommt Blaser (2019, S. 11) zu dem Schluss, dass sie zwar „einige Potenziale für den Bildungsbereich bereithalten. Jedoch deutet die Literatur auch an, dass ein Mehrwert auf den tatsächlichen Lernerfolg bislang nicht nachgewiesen werden konnte".

Entsprechend dieser Ausgangslage sollte die hier vorgestellte Studie nicht nur Aufschluss über konkrete Rezeptionsmuster geben, sondern auch über die Wirkung der Rezeption – konkret über den stattfindenden Wissenserwerb. Da Wissenserwerb – anders als Rezeptionsmuster – nicht einfach per Blickaufzeichnung beobachtet werden kann[3], ist eine Kombination mit weiteren Methoden notwendig (siehe auch Niemann und Scheuermann in diesem Band). Entsprechend kam bei der Untersuchung von KATRIN VR ein klassisch zu nennender Methodenmix zum Einsatz, bei dem die Blickaufzeichnung mit einem Wissenstest und der Methode des Lauten Denkens[4] kombiniert wurde.

[3] Durch den Einsatz der Methode der Blickaufzeichnung können jedoch – wie in Abschn. 5 konkret deutlich werden wird – Daten zur Aufmerksamkeitsverteilung gewonnen werden, die bei der Untersuchung von Wissenserwerb mindestens indikatorische Bedeutung haben.

[4] Bei der Methode des Lauten Denkens werden Versuchspersonen gebeten, spontan und unreflektiert zu verbalisieren, was ihnen bei der Nutzung – in diesem Fall der VR-Umgebung – durch den Kopf geht. Eine detaillierte Einführung in das Verfahren liefert Bilandzic (2017).

Was den Aspekt des Wissens betrifft, so spielt in der hier untersuchten VR-Umgebung nicht primär das vermittelte Faktenwissen[5] eine Rolle, sondern das Verständnis des physikalischen Experiments selbst, also das Strukturwissen. Um die Veränderung des Strukturwissens der Versuchspersonen durch die Nutzung der VR-Umgebung zu erfassen, wurde die Concept-Mapping-Methode eingesetzt. Dies ist „eine Methode zur grafischen Darstellung von Wissensstrukturen" (Gehl 2013, S. 104), bei der Versuchspersonen gebeten werden, die wesentlichen Begriffe eines Themenbereichs zu nennen bzw. aus vorgegebenen Begriffen auszuwählen und diese miteinander in Beziehung zu setzen (Bsp.: A „gehört zu" B). Wird die Methode mehrfach und zu unterschiedlichen Messzeitpunkten eingesetzt, lassen sich der Wissenszuwachs, die Veränderung der Wissensqualität oder auch Veränderungen der Struktur der Wissensnetzwerke von Proband:innengruppen bestimmen.

Mithilfe der Blickaufzeichnung und des Lauten Denkens wurden Daten zur Relevanz konkreter Umsetzungsmerkmale der VR-Umgebung beim Wissenserwerb erhoben. Die Aufzeichnung von Augenbewegungen während der Nutzung der VR-Umgebung ermöglicht es, die visuelle Aufmerksamkeitsverteilung der Lernenden direkt während des Rezeptionsprozesses zu erfassen und damit sehr konkret zu erkennen, in welcher Situation welche Merkmale der Umgebung im Fokus stehen. In Kombination mit der Methode des Lauten Denkens können Probleme bei der Nutzung, aber auch besonders positive Nutzungserlebnisse, gezielt einzelnen Merkmalen des jeweils rezipierten Produkts zugeordnet werden.

Aufgrund der Corona-Pandemie konnten die Proband:innen auf zwei Wegen teilnehmen: entweder vor Ort im Rezeptionslabor oder online. Bei Letzterem musste aus technischen Gründen auf die Blickaufzeichnung verzichtet werden (vgl. Abb. 3).

4.1 Ablauf

Zu Beginn der Untersuchung wurden die Teilnehmer:innen gebeten, eine Concept Map zum KATRIN-Experiment zu erstellen. Dafür hatten sie maximal 15 min Zeit und konnten aus einem Set an vorgegebenen Begriffen und Beziehungen wählen.

Danach absolvierten sie die geführte Tour in der KATRIN VR-Umgebung über die normale Browseroberfläche, d. h. ohne Virtual-Reality-Brille, da dies das

[5] Faktenwissen ist „Wissen über einzelne, voneinander unabhängige Informationseinheiten" (Gehl 2013, S. 72), welches isoliert und ohne weitere Kenntnisse des Kontexts abgerufen werden kann. Im hier vorliegenden Fall also beispielsweise das Wissen um den Namen des Experiments oder um seinen Standort.

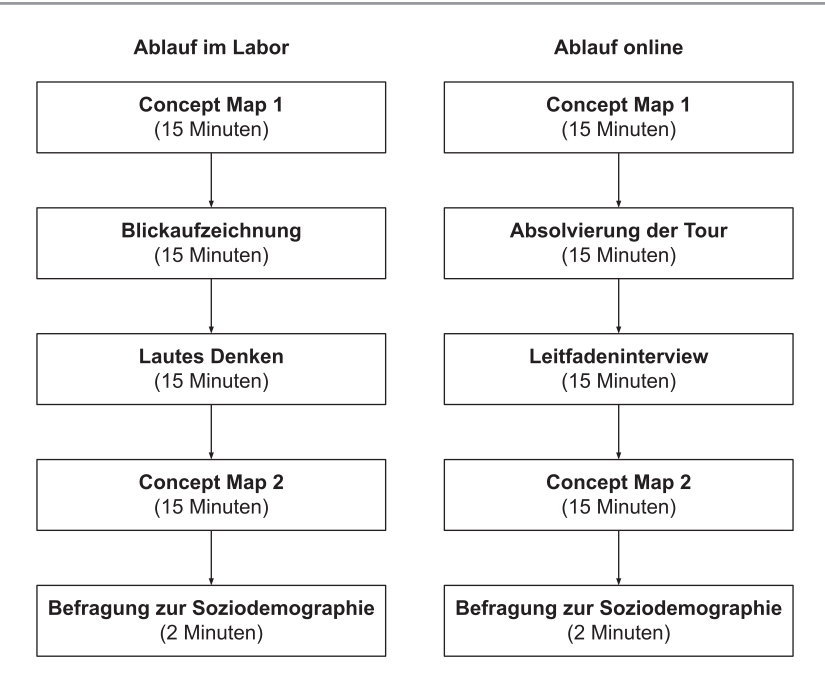


Abb. 3 Ablauf der Untersuchung in den beiden Szenarien Labor und online.

Szenario mit dem deutlich größeren Nutzungspotenzial unter Studierenden darstellt. Im Labor wurden währenddessen die Blicke der Proband:innen mithilfe eines Eye-Tracking-Geräts (SMI iView X) aufgezeichnet.

Nach Abschluss der Tour wurde den im Labor teilnehmenden Personen ein Video der gerade erfolgten Blickaufzeichnung vorgespielt. Während des Abspielens sollten sie laut darüber nachdenken, was sie getan haben und was ihnen dabei aufgefallen ist.

Danach folgte die zweite Concept Map, bei der aus dem gleichen Begriffs- und Beziehungsreportoire wie zuvor erneut das KATRIN-Experiment beschrieben werden sollte. Zum Abschluss wurde eine Reihe von Fragen zur Soziodemografie und zu vorherigen Kontakten mit dem KATRIN-Experiment gestellt.

Das Online-Verfahren unterschied sich aufgrund technischer Beschränkungen geringfügig vom Laborsetting. Die Teilnehmer:innen erstellten ebenfalls eine Concept Map vor und nach der Nutzung der VR-Umgebung und nahmen an der geführten Tour in der KATRIN-VR-Umgebung über die Software Zoom teil. Allerdings wurde keine Blickaufzeichnung durchgeführt und es wurde ihnen ent-

sprechend im Anschluss keine Aufzeichnung der Tour vorgespielt. Stattdessen kam ein Leitfadeninterview zum Einsatz, um Erfahrungen und Eindrücke zu sammeln.

4.2 Daten

Insgesamt nahmen zehn Personen an der Studie teil, sieben in der Laborumgebung und drei online. Bei einer Person wurde das Laute Denken aufgrund eines technischen Defekts nicht aufgezeichnet. Bei drei Personen (eine im Online-Setting, zwei im Laborsetting) hatte die VR-Tour kurz vor Ende einen technischen Defekt und wurde daraufhin abgebrochen. Dennoch wurden in diesen Fällen die beiden Concept Maps sowie das Leitfadeninterview bzw. das Laute Denken durchgeführt.

Das Durchschnittsalter der Proband:innen lag bei 22 Jahren. Der jüngste Befragte war 20 Jahre alt, der älteste war 36 Jahre alt. Acht von ihnen studierten zum Zeitpunkt der Befragung in einem Bachelorstudiengang Physik, fünf befanden sich im sechsten Semester und die übrigen drei im zweiten, achten bzw. zehnten Semester. Ein Befragter studierte Meteorologie (B.A.) im sechsten Semester und ein weiterer hatte bereits vor sieben Jahren eine Promotion in Physik abgeschlossen. Acht Teilnehmer waren männlich, zwei Teilnehmerinnen weiblich.

Ihre Vorerfahrungen mit dem KATRIN-Experiment waren sehr unterschiedlich. Während einige das Experiment besucht bzw. in einem Fall dort gearbeitet hatten, hatten die meisten Teilnehmer:innen nur in einer Vorlesungssitzung davon gehört. Man kann die Proband:innen somit zwar nicht als Lai:innen bezeichnen, die Mehrheit jedoch auch nicht als Expert:innen bezüglich KATRIN. Stattdessen können sie als Semi-Expert:innen angesehen werden, die mit einem gewissen Hintergrundwissen, aber ohne Detailwissen, die Umgebung nutzten.

5 Auswertung der Forschungsdaten

5.1 Allgemeine Beurteilung durch Proband:innen

Alle Teilnehmer:innen äußerten sich zufrieden mit der virtuellen Umgebung des KATRIN-Experiments und hielten sie für eine wertvolle und zugleich unterhaltsame Erfahrung, die sich für den Einsatz im universitären Unterricht eigne.

Besonders gelobt wurden die Möglichkeit, sich in der 360°-Umgebung frei umzusehen, die animierten Erklärfilme sowie die interaktiven Teile mit

dem Röntgenblick und die gute Orientierung innerhalb der Tour durch die eingeblendete Navigationsübersicht. Auch die humorvollen Elemente, wie z. B. die Darstellung einer Eistüte im Magen des Guides beim Einschalten des Röntgenblicks, wurden geschätzt.

Kritik äußerten nur einzelne Proband:innen. Ein Teilnehmer erwähnte, dass der Ablauf der Tour nicht mit dem realen KATRIN-Experiment übereinstimme. Ein anderer war sich über die Bedeutung eines bestimmten Begriffs unsicher, allerdings handelte es sich dabei um den Teilnehmer mit dem größten Fachwissen über KATRIN. Zwei Studierende erwarteten, dass es am Ende der Führung eine Art Datenanalyse des KATRIN-Experiments geben würde. Zwei weitere gaben an, dass sie die Interaktion mit dem Spektrometer nicht sofort verstanden hätten.

Häufiger beklagten sich die Teilnehmer:innen über technische Probleme, wie Verzögerungen bei den interaktiven Teilen – was es schwierig machte, die durch die Interaktion verursachten Unterschiede zu erkennen – oder Abbrüche während der Tour. Diese Probleme sind jedoch auf den spezifischen, technischen Aufbau der Studie statt auf die VR-Umgebung selbst zurückzuführen.

5.2 Wissenserwerb

Mit einer Ausnahme gaben alle Teilnehmenden an, dass sie während der Führung etwas gelernt hätten und das KATRIN-Experiment danach besser verstünden.

Diese Selbsteinschätzungen werden durch die Ergebnisse der vor und nach der Tour erstellten Concept Maps bestätigt. Um die Korrektheit zu bewerten, wurden die Maps mit einer Referenzkarte (vgl. Abb. 4) verglichen, die von dem Physiker erstellt wurde, der auch als virtueller Guide in der VR-Umgebung fungiert.

Die Zahl der richtigen Propositionen stieg von durchschnittlich 10 richtigen Propositionen vor der Tour auf 13,6 nach der Tour. Gleichzeitig sank die durchschnittliche Anzahl falscher Aussagen um 1,3, von 2,2 auf 0,9. Nur ein Teilnehmer hatte in der zweiten Map eine geringere Anzahl richtiger Aussagen und eine höhere Anzahl falscher Aussagen. Ein derartiger Anstieg der korrekten Aussagen wurde auch bei einem Laienpublikum für andere virtuelle Präsentationsformen zum KATRIN-Experiment beobachtet (Klein et al. 2021, S. 7).

Darüber hinaus gibt es weitere Indikatoren, die dabei helfen können, Veränderungen im Wissensstand zu bewerten, etwa die Netzwerkdichte oder die Zentralität von Begriffen in den Concept Maps. Wegen des Schwerpunkts dieses Artikels auf physiologischen Messungen werden die Studienergebnisse zum Aspekt Wissenserwerb an dieser Stelle aber nicht detaillierter ausgeführt.

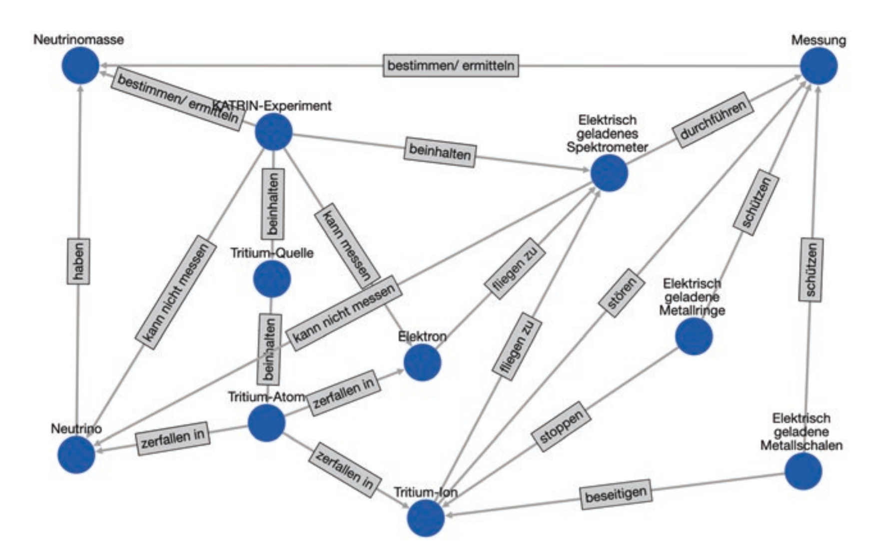


Abb. 4 Referenzkarte, mit der die Concept Maps der Teilnehmer:innen verglichen wurden.

5.3 Aufmerksamkeitsverteilung

Um den *Grad der Aufmerksamkeit und des Interesses* während der Rezeption zu analysieren, wurde zunächst erfasst, wie lange die Laborteilnehmer:innen zentrale Elemente der virtuellen Umgebung mit ihren Augen fixierten. Diese zentralen Elemente (sogenannte Areas of Interest, kurz AOI) waren der Guide, das Navigationspanel, die aktivierte Röntgenansicht und die dazugehörige Legende, die animierten Erklärvideos sowie die Bedienfelder für die eigenständige Interaktion mit dem Experiment (vgl. Abb. 5).

Die meiste Zeit – im Schnitt drei Minuten und 41 s – war der Blick der Teilnehmer:innen auf den Guide gerichtet, er fungiert somit als zentraler Ankerpunkt der Tour. Etwas mehr als drei Minuten verbrachten die Proband:innen damit, sich „umzuschauen“, d. h. mit der Maus den Ausschnitt des Raumes zu bewegen und so die Möglichkeiten der 360°-Umgebung zu nutzen.

Die anderen kodierten Elemente fanden hinsichtlich ihrer absoluten Betrachtungszeit weniger Beachtung (vgl. Abb. 6). Da sie jedoch nicht während der gesamten Tour zu sehen sind, ist in diesem Fall der Anteil an der Betrachtungszeit während ihrer tatsächlichen Sichtbarkeit relevanter. Auf diesen wird im Folgenden eingegangen.



Abb. 5 Beispielhafte Areas of Interest (AOIs) in der VR-Umgebung. (Quelle: eigene Darstellung)

Guide 03:41
Umschauen 03:10
Röntgenblick (aktiviert) 02:17
Bedienfeld (Hochspannungsspektrometer) 00:46
Animation (Messprinzip II) 00:45
Navigationsübersicht 00:26
Animation (Messprinzip I) 00:23
Bedienfeld (Tritiumzerfall) 00:16
Legende zur Tritiumquelle 00:11
Bedienfeld (Spannung Elektroden) 00:07
Button 'Weiter' 00:06
0 50000 100000 150000 200000
Millisekunden

Abb. 6 Durchschnittliche Dauer der summierten Fixationen.

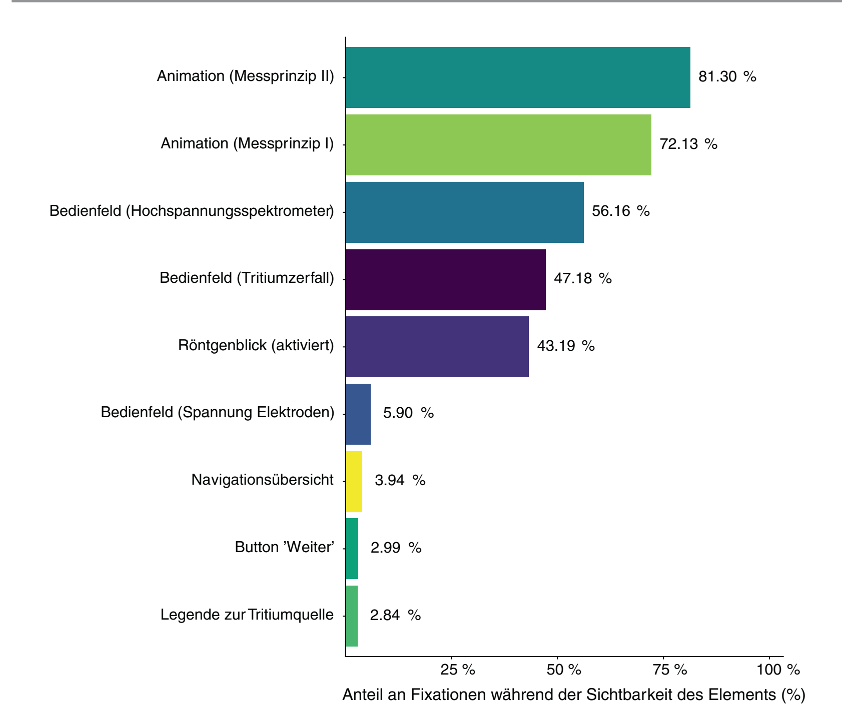


Abb. 7 Anteil an Fixationen während der Sichtbarkeit.

Abb. 7 zeigt das Verhältnis aus zeitlicher Sichtbarkeit und Betrachtungszeit: Beispielsweise wurde der animierte Erklärfilm „Messprinzip II" während seiner Sichtbarkeit von den Teilnehmer:innen im Durchschnitt 81,3 % der Zeit fixiert. Eine ähnlich hohe Quote kann auch der erste animierte Erklärfilm der Tour aufweisen (72 %). Diese hohe Aufmerksamkeit, die beiden animierten Erklärfilmen zuteilwurde, zeigt deren Bedeutung während der Tour.

Doch ist es nicht nur interessant zu analysieren, wie lange zentrale Elemente der VR-Umgebung angeschaut wurden, sondern auch in welcher Reihenfolge dies passierte (Stichworte *Selektions-* und *Erschließungsstrategien*).

Abb. 8 stellt die Betrachtungsreihenfolge der AOIs für jede Teilnehmer:in im Labor einzeln dar. Die dargestellten Muster der Sequence Charts[6] sind insgesamt

[6] Eine Sequence Chart visualisiert die Reihenfolge, in der AOIs betrachtet wurden, indem auf der x-Achse der zeitliche Ablauf der Blickaufzeichnung und auf der y-Achse die einzelnen AOIs aufgetragen werden. In Abb. 8 kennzeichnet ein grauer Balken, dass ein Element zwar sichtbar ist, aber nicht angeschaut wird, während eine Einfärbung Fixationen kennzeichnet.



Abb. 8 Sequence Charts aller Teilnehmer:innen im Labor. Die Abbildung zeigt die Abfolge, in der verschiedene AOIs während der Tour angeschaut wurden (AOI-Anordnung identisch bei allen Teilnehmer:innen). Ausgeschlossen ist die AOI des Tourguides, da dort die Blickdaten aus technischen Gründen mit einer anderen Software kodiert werden mussten.

relativ ähnlich[7], was darauf hinweist, dass auch die Aufmerksamkeitsverteilung während des Rundgangs – gemessen über den Blick – ähnlich ist und damit auch die Rezeption der Tour selbst.

6 Fazit

In der beschriebenen Studie wurde eine VR-Umgebung zu einem physikalischen Großexperiment, dem sogenannten Karlsruhe Tritium Neutrino Experiment (KATRIN), mithilfe eines Methodenmixes aus Blickaufzeichnung, Concept Mapping und Befragung untersucht.

KATRIN VR gibt den Nutzer:innen die Möglichkeit, über eine geführte Tour das Experiment, seine Abläufe und Ziele interaktiv kennenzulernen. Diese relativ neue mediale Form kann im konkreten Fall beispielsweise eingesetzt werden, um Studierenden in physikalischen Studiengängen wissenschaftliches Wissen zum Experiment zu vermitteln. Die Evaluation dieses Einsatzzwecks – die Nutzung durch Studierende mit dem Ziel der Wissensvermittlung – stand hier im Fokus.

Die Ergebnisse zeigen, dass sich das Strukturwissen der Proband:innen nach der Nutzung der KATRIN VR Umgebung positiv verändert hat: Die Teilnehmer:innen erlangten ein besseres Wissen und Verständnis für das KATRIN-Experiment. Dies gaben die Befragten nicht nur in Selbstauskünften an, sondern auch die erstellten Concept Maps belegen dies durch einen deutlichen Zuwachs korrekter Aussagen über das Experiment im Vorher-Nachher-Vergleich.

Darüber hinaus zeigen die Daten der Blickaufzeichnung, dass die Aufmerksamkeit bei allen Teilnehmer:innen ähnlich verteilt war und sich auf die zentralen Elemente der Informationsverbreitung konzentrierte. Besondere intensiv wurden dabei die beiden während der Tour abgespielten animierten Erklärfilme sowie die interaktiven Elemente – mit denen Veränderungen im Experiment durchgeführt werden konnten – betrachtet. Aus einer produktanalytischen Perspektive liegt somit nahe, dass die intendierte Schwerpunktsetzung der Produzent:innen in der Praxis grundsätzlich funktioniert und sich der Wissenszuwachs der Proband:innen nicht zuletzt aus diesen Elementen ergibt.

Zudem bewerteten die Teilnehmer:innen während des Lauten Denkens bzw. während der Leitfadeninterviews die VR-Umgebung insgesamt positiv. Ins-

[7] Die Lücke in den Daten des letzten Teilnehmers ist auf ein technisches Problem während des Experiments zurückzuführen.

besondere die Verständlichkeit der Tour im Allgemeinen, die interaktiven Elemente und die animierten Erklärfilme wurden gelobt. Demgegenüber wurden nur wenige negative Punkte genannt, die sich zudem mehrheitlich auf technische, durch das Untersuchungssetting verursachte Probleme – etwa Performanceeinbrüche bei der Darstellung – bezogen.

Basierend auf diesen Ergebnissen scheinen VR-Umgebungen für den Einsatz in der universitären Ausbildung gut geeignet zu sein und könnten traditionelle Lehrformen – gerade vor dem Hintergrund gänzlich digitaler oder hybrider Lehre – ergänzen.

6.1 Einschränkungen und Methodenreflexion

Da es sich bei der hier dargestellten Studie um eine stark qualitativ ausgerichtete Untersuchung mit einer kleinen Teilnehmer:innenzahl und nur einer getesteten VR-Umgebung handelt, muss die Verallgemeinerbarkeit der Ergebnisse zurückhaltend bewertet werden. Kontextfaktoren – z. B. Studienfach, Alter, Thema und konkrete Gestaltung der VR-Umgebung usw. – könnten das Ergebnis beeinflussen.

Dieses Problem stellt sich bei der Verwendung der Methode der Blickaufzeichnung zur Analyse von Rezeptionsprozesses realer Medienprodukte generell. Aufgrund der relativ aufwendigen Durchführung der Erhebungen und der ebenso zeitintensiven Auswertung, die sich bislang kaum automatisieren lässt, können in den meisten Fällen nur geringe Stichprobengrößen untersucht werden.

Ebenso ist es mit Blickaufzeichnungsstudien und den hier benutzten ergänzenden Methoden nur möglich, die Rezeption direkt sowie unmittelbare Wirkungen derselben zu erfassen. Aussagen über mittel- und langfristige Medienwirkungen hingegen lassen sich so nicht treffen.

Zudem muss darauf hingewiesen werden, dass zwar die aufgezeichneten Blickbewegungen quasi unmittelbare Reaktionen auf einen Reiz – in diesem Fall die KATRIN VR Umgebung – sind, für deren Kontextualisierung (Warum wurde hier hin oder dahin geschaut?) aber Selbstaussagen der Proband:innen notwendig sind, deren Wahrheitsgehalt etwa durch Effekte der sozialen Erwünschtheit beeinträchtigt sein können.

Insgesamt macht das Beispiel für die Anwendung einer physiologischen Messung mit der Methode der Blickaufzeichnung deutlich, dass deren praxistaugliche Verwendung im wissenschaftskommunikativen Evaluationskontext nur in einem Methodenmix möglich ist. Wenn jedoch die fachlichen, zeitlichen und finanziellen Möglichkeiten zur Verfügung stehen, bietet diese Methode im

Bereich der Medienrezeption eine Möglichkeit, die mit anderen Verfahren so nicht erreicht werden kann: Den direkten, kaum manipulierbaren Einblick in den Kommunikationsprozess zwischen Nutzer:innen und medialem Artefakt.

Literatur

Bilandzic H (2017) Lautes Denken. In: Mikos L, Wegener C (Hrsg) Qualitative Medienforschung: Ein Handbuch, 2. Aufl. UVK, Konstanz, S 406–413

Blaser N (2019) Identifizierung von Merkmalen wissenschaftlicher 360°-Videos: Literaturüberblick und vergleichende Videoanalyse. In: Science In Presentations Arbeitsberichte, #8. https://doi.org/10.5445/IR/1000132276

Geise S (2011) Eyetracking in der Kommunikations- und Medienwissenschaft: Theorie, Methode und kritische Reflexion. SCM Studies in Communication and Media (2):149–263. https://doi.org/10.5771/2192-4007-2011-2

Gehl D (2013) Vom Betrachten zum Verstehen. Springer Fachmedien, Wiesbaden. https://doi.org/10.1007/978-3-531-19823-1

Klein M, Humm C, Köllenberger L, Niemann P, Scheuermann Y, Schrögel P, Valerius K (2021) Virtual tours to the KATRIN experiment. Proceedings of 37th International Cosmic Ray Conference — PoS ICRC2021:1376. https://doi.org/10.22323/1.395.1376

Christian Humm, M.A., ist Fachreferent im Büro des Universitätspräsidenten an der Universität des Saarlandes. Zuvor hat er am Karlsruher Institut für Technologie (KIT) im Department Wissenschaftskommunikation geforscht und gelehrt. Er studierte Medien- und Politikwissenschaft an der Universität Trier und der Lancaster University.

Philipp Niemann ist stellvertretender Direktor und wissenschaftlicher Leiter des Nationalen Instituts für Wissenschaftskommunikation (NaWik). Zuvor war er als Nachwuchsgruppenleiter im Department für Wissenschaftskommunikation am Karlsruher Institut für Technologie (KIT) tätig. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftskommunikation, qualitative Rezeptionsforschung und politische Kommunikation.



Grundlagenbeitrag: Inhaltsanalysen inklusive Medienanalysen

Sabrina H. Kessler und Nina Wicke

Zusammenfassung

Der Grundlagenbeitrag fokussiert auf die Methode der Inhaltsanalyse inkl. Medienanalyse und reflektiert den Einsatz dieser im Bereich der Evaluation von Wissenschaftskommunikation. Nachdem einleitend deren Relevanz reflektiert wird, wird diese vorgestellt, indem Untersuchungsgegenstände, Analyseprozesse und Ziele der quantitativen/standardisierten und qualitativen Inhaltsanalyse erläutert werden. Herausgearbeitet werden dann die Analyseschwerpunkte der Inhalts- und Medienanalysen im Bereich der Evaluation von Wissenschaftskommunikation und ihrer Begleitforschung. Diese sind u. a. (1) Modi der Wissenschaftskommunikation, (2) die Genauigkeit der Berichterstattung, (3) die Darstellung, das Framing und die Bewertung von Wissenschaft und wissenschaftlichen Erkenntnissen, und (4) Dialogizität und Funktionalität der Wissenschaftskommunikation bzw. Funktionen der Öffentlichkeitsarbeit. Im letzten Punkt wird ein Ausblick gegeben und relevante Forschungslücken werden herausgestellt.

S. H. Kessler (✉)
Institut für Kommunikationswissenschaft und Medienforschung der Universität Zürich, Universität Zürich, Zürich, Schweiz
E-Mail: s.kessler@ikmz.uzh.ch

N. Wicke
Institut für Kommunikationswissenschaft der Technischen Universität Braunschweig, Braunschweig, Deutschland
E-Mail: nina.wicke@gmx.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_14

1 Einleitung

Bei der Evaluation des Inhaltes von Wissenschaftskommunikation können alle Formen der Kommunikation analysiert und bewertet werden. Das heißt Wissenschaftskommunikation, die sich auf wissenschaftliches Wissen und wissenschaftliche Arbeit innerhalb und außerhalb der institutionalisierten Wissenschaft konzentriert, ist Untersuchungsgegenstand (vgl. Schäfer et al. 2015, S. 13). Dabei kann die Evaluation Kommunikationsinhalte von verschiedenen Kommunikator:innen betreffen, welche mit der Kommunikation jeweils unterschiedliche Ziele verfolgen können. Die wichtigsten Wissenschaftskommunikator:innen sind Wissenschaftler:innen selbst, (Wissenschafts-)Journalist:innen sowie die Kommunikationsabteilungen wissenschaftlicher Einrichtungen (Wicke 2022). Darüber hinaus gibt es weitere Wissenschaftskommunikator:innen, wie Nichtregierungsorganisationen (NGOs), aktivistische Organisationen, Think Tanks oder an Wissenschaft interessierte Laien (Fähnrich 2018a, 2018b).

Wissenschaftskommunikation vermittelt das Verständnis von Wissenschaft in der Gesellschaft. Dabei erfüllt sie grundlegende Funktionen wie Information und Wissensvermittlung zu Forschung und Wissenschaft, die Legitimation der gesellschaftlich zur Verfügung gestellten (finanziellen) Ressourcen und bietet Bürger:innen ein Partizipationsforum für relevante Forschungsthemen (Bubela et al. 2009; Burns et al. 2003; Pfenning 2012). Ihre Ziele reichen von der Förderung eines größeren öffentlichen Verständnisses für und Engagement mit Wissenschaft, über die Schaffung von Akzeptanz und Vertrauen in wissenschaftliche Erkenntnisse, bis hin zum Reputationsmanagement von Wissenschaftler:innen und wissenschaftlichen Institutionen (Bubela et al. 2009; Burns et al. 2003). Die Evaluation von Wissenschaftskommunikation kann sich mitunter an der Erreichung dieser Ziele und ihrer Funktionalität orientieren.

Die Evaluation der öffentlichen Kommunikation von wissenschaftlichen Inhalten ist von großer Bedeutung, da sich deren Darstellung auf die Wahrnehmung, das Verständnis, die Einstellungen und auf das Vertrauen der Rezipient:innen in Bezug auf Wissenschaft und wissenschaftliche Themen auswirkt (Schäfer et al. 2019). Inhaltsbezogene Evaluationsmethoden können hier u. a. eine Systematisierung der vielfältigen Kommunikationsansätze nach deren Effizienz und Effektivität leisten und die Auswahl bestmöglicher Praktiken ermöglichen, um wissenschaftliche Inhalte optimal vermitteln zu können (Pfenning 2012; Ziegler et al. 2021).

Die Inhaltsanalyse und insbesondere die manuelle, standardisierte Inhaltsanalyse gilt als zentrale Methode der Kommunikations- und Wissenschaftskommunikationsforschung (Kessler und Schäfer 2022; Kessler et al. 2020). Erste inhaltsanalytische Analysen der Wissenschaftskommunikation entstanden in den späten 1960er Jahren an der Schnittstelle von Wissenschaftspädagogik, Wissenschaftssoziologie, Massenkommunikation und Museologie (Schäfer et al. 2019). Im Bereich der Evaluation von Wissenschaftskommunikation ist sie eine der am häufigsten angewendeten Methoden, sei es bei Fragen der Qualität, Akkuratheit oder Verzerrung von Kommunikation zu wissenschaftlichen Themen (Kessler und Schäfer 2022). Die Wichtigkeit und der häufige Einsatz wurden sowohl in Analysen von Tagungsbeiträgen als auch in Metaanalysen von Studien zur Wissenschaftskommunikation nachgewiesen (Kessler et al. 2020; Schäfer 2012).

2 Einsatz der Inhalts- und Medienanalyse im Bereich der Evaluation von Wissenschaftskommunikation

Die Methode der Inhaltsanalyse wurde erstmals in Berelsons soziologischem Werk von 1952 theoretisch fundiert vorgestellt: „Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication“ (Berelson 1952, S. 18). Heute unterscheiden sich die Methodendefinitionen in einigen Punkten von Berelsons erstem Vorschlag: Nach Früh (2017) ist die Inhaltsanalyse eine empirische „Methode zur systematischen, intersubjektiv nachvollziehbaren Beschreibung inhaltlicher und formaler Merkmale von Mitteilungen, meist mit dem Ziel einer darauf gestützten interpretativen Inferenz auf mitteilungsexterne Sachverhalte“ (S. 29). Methodenhandbücher, welche die Methode eingehend beschreiben und reflektieren, sind bspw. von Brosius et al. (2016), Früh (2017), Rössler (2017) sowie Scheufele und Engelmann (2009).

Die Untersuchungsgegenstände von Inhaltsanalysen sind in der Regel Medienprodukte. Dies können Artikel aus Tageszeitungen sein, Rundfunksendungen, Filme, Werbespots, aber auch Twitter-, Instagram- und Facebook-Postings, Nutzer:innenkommentare unter Online-Beiträgen etc. Sämtliche Formen von textlichen oder visuellen Inhalten, die öffentlich kommuniziert werden, können analysiert werden.

Inhaltsanalysen lassen sich nach ihrem Standardisierungsgrad unterscheiden in qualitative Inhaltsanalysen, welche dezidiert die Individualität einzelner spezifisch ausgewählter Medienangebote berücksichtigen, und quantitative Inhaltsana-

lysen, welche quantifizierend bzw. messend vorgehen und meistens auf größere Stichproben angewendet werden (Rössler 2017; Scheufele und Engelmann 2009). Die quantitative/standardisierte Inhaltsanalyse reduziert die Komplexität der Berichterstattung, indem sie formale und inhaltliche Merkmale großer Textmengen erfasst und reduktiv analysiert, d. h. auch deren zentrale Muster herausarbeitet (Brosius et al. 2016; Rössler 2017). Bei der manuellen Inhaltsanalyse wird die Codierung, im Gegensatz zur automatisierten Inhaltsanalyse, nicht durch einen Computer durchgeführt, sondern durch menschliche Codierer:innen. Die automatisierte (computergestützte) Inhaltsanalyse hat in der Kommunikationswissenschaft in den vergangenen Jahren zwar an Bedeutung gewonnen (Rössler 2017; Wirth et al. 2015), Studien im Bereich der Evaluation von Wissenschaftskommunikation lassen sich allerdings bisher nur vereinzelt finden.

Die klassische manuelle, standardisierte Inhaltsanalyse lässt sich in der Durchführung in vier Phasen einteilen: Planungs-, Entwicklungs-, Anwendungs- und Auswertungsphase. Sobald Forschungsfragen und Untersuchungsgegenstände definiert sind, legen die Forschenden in der ersten Phase der Planung fest, welche (medialen) Inhalte (Analyse- und Codiereinheiten) sie untersuchen wollen und wie gegebenenfalls eine Stichprobe aus der Grundgesamtheit gezogen wird (Früh 2017; Rössler 2017). Die Ausarbeitung des Untersuchungsinstrumentes, d. h. des Codebuchs, ist in der Entwicklungsphase zentral. Dies ist die Phase der Operationalisierung, d. h. des Messbarmachens der relevanten Konstrukte. Um intersubjektive Nachvollziehbarkeit zu gewährleisten, muss das Codebuch umfassende Definitionen der formalen und inhaltlichen Variablen und Ausprägungen, genaue Codieranweisungen und auch Codierbeispiele enthalten (Brosius et al. 2016). Die Variablenauswahl und -bildung wird in der standardisierten Inhaltsanalyse theoriegeleitet vorgenommen. Die Variablen werden kontextabhängig und induktiv interpretiert und codierbare Ausprägungen erstellt (Kessler et al. 2022). Das so entwickelte Codebuch wird in der Testphase umfangreichen Prüfungen im Hinblick auf Vollständigkeit, Reliabilität und Anwendbarkeit unterzogen. Bevor die Phase der Anwendung, die Hauptcodierung, beginnt, findet eine systematische und umfangreiche Reliabilitätsüberprüfung statt. Hier wird die Inter- und/oder Intracoderreliabilität gemessen und kontrolliert, ob die Codierer:innen ein gemeinsames inhaltliches Verständnis der Variablen haben und jeweils das Gleiche codieren bei demselben Untersuchungsmaterial. Die Inhaltsanalyse endet mit der Phase der (deskriptiven und statistischen) Auswertung der Daten inklusive der Darstellung und Interpretation der Ergebnisse. Ziel der Inhaltsanalyse ist, neben der Beschreibung der Inhalte auch darüber hinaus Schlussfolgerungen (Inferenzen) bspw. hinsichtlich des historischen, politischen oder sozialen Kontextes, der Kommunikator:innen oder

Rezipient:innen zu ermöglichen. Ein Vorteil der Inhaltsanalyse ist, dass sie ein zeitunabhängiges, nicht-reaktives Verfahren ist, d. h. ihr Untersuchungsgegenstand verändert sich nicht. Sie ist so beliebig reproduzierbar und modifizierbar. Wichtigste Gütekriterien, um die Qualität der Inhaltsanalyse, d. h. der Messung und des Untersuchungsinstrumentes zu kontrollieren, sind die der *Reliabilität* und *Validität* (Brosius et al. 2016; Rössler 2017; Scheufele und Engelmann 2009). Die Reliabilität betrifft die Zuverlässigkeit der Messung bei wiederholter Anwendung des Messinstrumentes. Die Validität betrifft die Gültigkeit der Messung und gibt an, ob das Instrument tatsächlich auch das misst, was es messen soll.

Qualitative und quantitative Inhaltsanalysen werden im Feld der Wissenschaftskommunikationsforschung gleichermaßen eingesetzt (Kessler et al. 2020; Schäfer 2012). Die Studien untersuchen häufig die Darstellung von wissenschaftlichen Themen in einzelnen Medien und Ländern mit einer klaren Ausrichtung auf westliche Länder und deren Printmedien (für einen Überblick siehe Schäfer 2012). In letzter Zeit werden Analysen von wissenschaftsbezogenen Medieninhalten traditioneller Medien zunehmend durch Studien zur Online- und Social-Media-Kommunikation ergänzt (Kessler und Schäfer 2022; Wicke 2022). Die meisten Studien sind Ein-Disziplin-Analysen, die hauptsächlich (mehr als 80 %) Naturwissenschaften oder verwandte Forschungsfelder als Untersuchungsgegenstände haben (Kessler und Schäfer 2022; Schäfer 2012).

Ein großer Teil der Studien, welche Inhaltsanalysen nutzen, baut auf normativen Annahmen zur Rolle der Massenmedien und Kommunikator:innen in der Gesellschaft auf. Vor diesem Hintergrund beschäftigen sich viele Inhaltsanalysen mit Berichterstattungsstilen oder -qualität und untersuchen, wie effektiv und adäquat Kommunikationsangebote Bürger:innen mit wissenschaftsbezogenen Informationen versorgen, um eine Meinungsbildung in der demokratischen Gesellschaft zu ermöglichen (Kessler et al. 2022). Themenspezifische Studien in der Wissenschaftskommunikationsforschung fokussieren entsprechend dabei – noch mehr als Studien in anderen thematischen Bereichen – die Frage, wie qualitativ, evidenzbasiert, akkurat bzw. verzerrt die wissenschaftlichen Themen, Ereignisse oder Befunde dargestellt werden und welche Themenaspekte, Frames, Akteur:innen, oder Meinungen in der Berichterstattung vorkommen (Kessler und Schäfer 2022). Darunter fallen bspw. Studien, die einen Vergleich von In- und Output zwischen wissenschaftlichen Artikeln, Hochschul-Pressemitteilungen und Nachrichtenbeiträgen vornehmen (Heyl et al. 2020; Sumner et al. 2014, 2016). Weitere Analysegegenstände inhaltsanalytischer Untersuchungen im Feld entstammen der digitalen institutionellen Kommunikation, wie Facebook- und Twitter-Posts sowie Websites von Hochschulen (Bélanger et al. 2013; Metag und Schäfer 2019; Zhang und O'Halloran 2013). In jüngster Zeit stützen sich

Studien zudem auf computergestützte Inhaltsanalysen. So führten bspw. Walter et al. (2019) im Rahmen einer Twitter-Netzwerkanalyse eine automatisierte Inhaltsanalyse von Tweets von Wissenschaftler:innen zum Klimawandeldiskurs durch. Im Vergleich zu den journalistischen und institutionellen Formen der Wissenschaftskommunikation liegt bislang aber noch wenig Forschung zu den Kommunikationsinhalten von alternativen Wissenschaftskommunikator:innen vor (Fähnrich 2018b; Kessler et al. 2020; Schäfer et al. 2019).

3 Analyseschwerpunkte der Inhalts- und Medienanalysen im Bereich der Evaluation von Wissenschaftskommunikation und ihrer Begleitforschung

Inhaltsanalytische Studien von Wissenschaftskommunikation, die auch zu deren Einordnung und Evaluation herangezogen werden, weisen eine große Vielfalt an Analysepunkten bzw. Forschungsfoki auf. Häufige Analyseschwerpunkte sind:

1. Studien, die verschiedene *Modi der Wissenschaftskommunikation* identifizieren und evaluieren: So wird bspw. erfasst, wie medialisiert oder popularisiert die Berichterstattung ist (Peters et al. 2013; Rödder und Schäfer 2010; Weingart 2012). Bei der *Medialisierung* orientiert sich die Wissenschaftsberichterstattung verstärkt an Logik und Normen des Mediensystems (d. h. auch weniger nach denen des Wissenschaftssystems). Daraus folgt, dass diese Berichterstattung oft durch gesellschaftspolitische oder kulturelle Ereignisse ausgelöst wird, sich weniger auf wissenschaftliche Quellen stützt und konfrontativer sowie konfliktreicher ist (Peters 1994; Schäfer 2009). *Popularisierung* ist dadurch gekennzeichnet, dass wissenschaftliche Informationen (meist sehr positiv) präsentiert werden, die von Wissenschaftler:innen oder JournalistInnen erklärt, aber nicht problematisiert oder kritisch hinterfragt werden (Schäfer 2009). Diese Art der Berichterstattung wird nicht nur danach beurteilt, wie akkurat, sondern auch wie verständlich Erkenntnisse aus Wissenschaft und Forschung vermittelt werden (Gerhards und Schäfer 2011; Kohring 1997). Dies gelingt bspw. Wissenschaftsblogs häufig gut: Blog-Autor:innen nähern sich ihren Themen in einem alltagsnahen Stil und wechseln Erklärungen mit persönlichen Meinungen sowie humorvollen Bemerkungen ab, sodass Leser:innen ohne wissenschaftlichen Hintergrund die Inhalte verstehen können sollen (Kouper 2010; Mahrt und Puschmann 2014).

2. Studien, die die *Genauigkeit der Berichterstattung* gemessen an wissenschaftlichen Standards evaluieren: Diese Studien versuchen, die Genauigkeit der Medienberichterstattung über Wissenschaft zu bewerten, indem sie diese mit wissenschaftlichen Publikationen oder Pressemitteilungen vergleichen. Studien, welche dies im Online-Kontext untersuchen, sind oft getrieben von der Annahme, dass der Mangel an Qualitätskontrolle und journalistischem Gatekeeping im Internet zu minderwertigen, fälschlichen Darstellungen wissenschaftlicher Themen führen könnte (Barr 2011; Cacciatore et al. 2012). Ein Schwerpunkt dieser Medieninhaltsanalysen ist die Untersuchung der Unsicherheitsdarstellung von wissenschaftlicher Evidenz bei verschiedenen strategischen Kommunikator:innen, in unterschiedlichen Medien und/oder in Bezug auf verschiedene wissenschaftliche Themen (Cacciatore et al. 2012; Dudo et al. 2011; Guenther et al. 2019; Kessler 2016; Mellor 2010; Stocking und Holstein 2009). Deren Ergebnisse zeigen, dass die Medienberichterstattung in der Regel bis zu einem gewissen Grad von wissenschaftlichen Publikationen abweicht, übertrieben und sensationalisiert ist (z. B. Knudsen 2005), Erkenntnisse vereinfacht (Brechman et al. 2009) und Unsicherheiten oft gar nicht oder unzureichend darstellt (Dudo et al. 2011; Guenther et al. 2019; Kessler 2016; Stocking und Holstein 2009).
   Inhaltsanalysen mit Fokus auf der Evaluation der externen Kommunikation wissenschaftlicher Einrichtungen ziehen zur Erfolgsbestimmung von Reputationsbemühungen in der Regel die Berichterstattung überregionaler Medien als Indikator heran (Friedrichsmeier et al. 2015). Eines der etabliertesten Evaluationstools ist diesbezüglich die Medienresonanzanalyse (Zerfaß und Volk 2019; Raupp und Vogelgesang 2009). Zudem wird häufig die Darstellung eines Forschungsthemas in wissenschaftlichen Journals mit den dazugehörigen Pressemitteilungen und mit der anschließenden journalistischen Berichterstattung verglichen (Brechman et al. 2009, 2011; Bubela und Caulfield 2004; Sumner et al. 2014, 2016; Winters et al. 2019; Yavchitz et al. 2012). Diese Inhaltsanalysen zeigen u. a., dass übertriebene Darstellungen in Medienberichten mit Übertreibungen in Pressemitteilungen zusammenhängen. Zudem werden wichtige Details der wissenschaftlichen Studien, wie Finanzierung und Studienlimitationen, bereits in den Pressemitteilungen der wissenschaftlichen Einrichtungen weitgehend nicht thematisiert (Brechman et al. 2011; Sumner et al. 2014; Winters et al. 2019).
3. Studien, die *die Darstellung, das Framing und die Bewertung von Wissenschaft und wissenschaftlichen Erkenntnissen* analysieren und evaluieren: Um zu ermitteln, wie wissenschaftliche Themen medial konstruiert werden und welche Aufmerksamkeit welchen Themenaspekten zuteilwird, werden

(meist quantitativ angelegte) Inhaltsanalysen im Print- (Bohr 2020; Lopera und Moreno 2014; Shea 2015; Vestergård und Nielsen 2016) und Online-Kontext (Erviti et al. 2020; Lörcher und Taddicken 2017; Taddicken et al. 2019) durchgeführt. Die Framing-Forschung hat gezeigt, dass in der Medienberichterstattung ein und dieselben Themen hierbei unterschiedlich gerahmt werden können, d. h. spezifische Aspekte eines Themas ausgewählt und hervorgehoben werden (z. B. Gerhards und Schäfer 2011 für Humangenom-Forschung; Kessler 2016 für Medizin; Nisbet et al. 2003 für Biotechnologie; Ruhrmann et al. 2015 für molekulare Medizin; Schäfer und O'Neill 2017 für Klimawandel; Taddicken et al. 2020 für autonomes Fahren). Die mediale Darstellung von Wissenschaft und wissenschaftlichen Themen wird zudem dahingehend analysiert, wie kritisch bzw. positiv sie ist (Schäfer 2009; Vestergård und Nielsen 2017) und auch inwiefern geschlechtliche und ethnische Diversität repräsentiert werden (Bal und Sharik 2019).

4. Der Paradigmenwechsel in der Diskussion um das Verhältnis von Wissenschaft und Öffentlichkeit – weg von der Annahme eines Informations- und Kompetenzdefizits bei den Bürger:innen, dem sogenannten „Defizitmodell", hin zu partizipativen Ansätzen im Sinne einer öffentlichen Auseinandersetzung mit der Wissenschaft (für einen Überblick, siehe Akin 2017; Schmid-Petri und Bürger 2019) – führte dazu, dass vermehrt die *Dialogizität und Funktionalität der Wissenschaftskommunikation* evaluiert wird. Neuere Studien konzentrieren sich auf die Online-Kommunikation von Wissenschaftler:innen und wissenschaftlichen Einrichtungen, die es ermöglicht, direkt mit der Öffentlichkeit zu kommunizieren. Inhaltsanalysen untersuchen, wie sich Wissenschaftler:innen auf Social Media engagieren und wie sie interagieren (Hara et al. 2019; Jahng und Lee 2018; Jünger und Fähnrich 2019; Walter et al. 2019). Dabei werden häufig die Inhalte und deren kommunikative Funktion analysiert, der Grad und die Arten des Engagements der Wissenschaftler:innen und ihre Beziehung zu Nutzer:innen. Die Ergebnisse deuten darauf hin, dass Wissenschaftler:innen vor allem einseitig kommunizieren und oftmals keinen Dialog mit der Öffentlichkeit herstellen (Jahng und Lee 2018; Jünger und Fähnrich 2019; Walter et al. 2019). Auch Wissenschaftsblogs werden von Wissenschaftler:innen tendenziell dazu genutzt, sich selbst zu positionieren (Mahrt und Puschmann 2014; Shema et al. 2012).

   Die Forschung zur Wissenschaftskommunikation wissenschaftlicher Einrichtungen untersucht in Bezug auf die Social-Media-Praktiken üblicherweise community-bezogene Aspekte wie die Anzahl der Likes und Freund:innen/Abonnent:innen sowie den Inhalt von Postings (Bélanger et al. 2013;

Linvill et al. 2012, 2015; Su et al. 2017). Die Interaktion mit Follower:innen wird häufig hinsichtlich ihres informativen und dialogischen Potenzials analysiert. Bisherige Befunde deuten darauf hin, dass auch von Institutionen neue Medienplattformen bisher eher zur Informationsverbreitung als zum Engagement mit Stakeholder:innen genutzt werden (Lee und VanDyke 2015; Lee et al. 2017).

Nicht nur die Wissenschaftskommunikation der wissenschaftlichen Einrichtungen, sondern auch die der Wissenschaftsmuseen und NGOs zielt darauf ab, in ihren Kommunikationsansätzen und -strategien nutzer:innen- und dialogzentrierter zu werden. Inhaltsanalytische Studien zeigen jedoch, dass die meisten Wissenschafts- und Naturkundemuseen auf Webseiten und soziale Medien in einer traditionellen, einseitig geprägten Informationsübermittlung ihre Exponate und Aktivitäten bewerben (Capriotti et al. 2016; Jarreau et al. 2019; Jensen 2013). Auch NGO-Webseiten werden meist nur für die Bildungs- und nicht für die Aktivierungskommunikation genutzt (Yang und Taylor 2010). Die *Funktionen der Öffentlichkeitsarbeit*, insbesondere von alternativen Wissenschaftskommunikator:innen, werden zunehmend im Hinblick auf Informationsvermittlung und Beziehungsaufbau evaluiert. So wird bspw. untersucht, inwiefern Organisationen ihre Webseiten als Instrument für Medienarbeit, Spender:innenbeziehungen und Beziehungen zu Freiwilligen nutzen (Jun 2011; Yeon et al. 2007). Weitere Dimensionen dieser Inhaltsanalysen sind Interaktionsmöglichkeiten, die Nützlichkeit der Informationen für die Mitglieder und Freiwilligen, für die Öffentlichkeit und die Medien, sowie die Steigerung der Besucher:innenzahlen, die Generierung von Gegenbesuchen und die Leitbilder (Taylor et al. 2001; Yang und Taylor 2010). Inhaltsanalytische Methoden werden auch angewendet, um die dialogischen Strategien der Organisationen bei der Nutzung von Social Media-Plattformen wie Twitter und Facebook und ihre Beziehungen zu Stakeholder zu analysieren (Cho et al. 2014; Waters und Jamal 2011; Waters et al. 2009). Zur Evaluation von organisatorischen Informationsstrategien und PR-Aktivitäten auf Facebook wurden dann mitunter Informationsverbreitung und Involvement (Waters et al. 2009) sowie die Anzahl von Likes, Shares und Kommentaren zu den Postings codiert (Cho et al. 2014). Die Studie von Castillo-Esparcia et al. (2015) evaluierte bspw. die Performance von Think Tanks in sozialen Medien und sieht dabei Verbesserungsbedarf in allen gemessenen Dimensionen, d. h. in Bezug auf Sichtbarkeit, Reichweite, Interaktivität und Engagement.

4 Ausblick und Forschungslücken

Inhaltsanalysen inkl. Medienanalysen sind eine etablierte Methode im Bereich der Evaluation von Wissenschaftskommunikation und sie werden dies auch bleiben. Zukünftige inhaltsanalytische Studien könnten sich aber noch mehr den bisher nur wenig systematisch erforschten Gegenständen zuwenden, wie z. B. der Evaluation der Wissenschaftskommunikation in nicht-westlichen Ländern und von Disziplinen jenseits der Naturwissenschaften (Kessler und Schäfer 2022).

Da die Organisationskommunikation von wissenschaftlichen Einrichtungen und alternativen Kommunikator:innen wie Umweltaktivist:innen zwar zunimmt (Schäfer und Fähnrich 2020), aber in der Forschung noch nicht viel Aufmerksamkeit erhält (Fähnrich 2018b; Schäfer und Fähnrich 2020; Schäfer et al. 2019), könnte auch dieser Forschungsbereich der strategischen Wissenschaftskommunikation differenzierter untersucht werden. Studien zu Hochschulkommunikation betrachten bisher oft nur spezifische Kommunikationskanäle wie deren Webseiten oder Social-Media-Auftritte und berücksichtigen nicht die zugrunde liegenden Kommunikationskonzepte und deren Ziele (Metag und Schäfer 2017), sodass kein umfassendes Bild gezeichnet werden kann. Es mangelt zudem an langfristigen Analysen, die Veränderungen in der Hochschulkommunikation identifizieren (Zhang und O'Halloran 2013). Oftmals konzentrieren sich die Evaluationen der Kommunikation von wissenschaftlichen Institutionen auf die praktische Anwendbarkeit ihrer Erkenntnisse und lassen theoretische Grundlagen vermissen. Sie können daher nur bedingt verallgemeinert werden (Metag und Schäfer 2019).

Eine zukünftige Herausforderung ist auch die Evaluation neuer Formen der Wissenschaftskommunikation wie z. B. Science Center, Science Slams, Wissenschaftsfestivals oder Wissenschaftspodcasts (Fähnrich 2017; Wicke 2021). Um deren Rezeption und Wirkung besser zu verstehen, z. B. hinsichtlich der öffentlichen Wahrnehmung von Wissenschaft oder einer Erhöhung der scientific literacy, sollten auch die innerhalb solcher Formate kommunizierten Inhalte inhaltsanalytisch betrachtet werden (Wicke 2022).

Methodisch ist es empfehlenswert, zukünftig Forschungsinstrumente, d. h. Codebücher, vermehrt open access zu teilen und replizierend anzuwenden (Kessler und Schäfer 2022).[1] Bislang haben die meisten Einzelstudien individuelle

[1] Hier sei auf die Database of Variables for Content Analysis (DOCA) hingewiesen, welche Operationalisierungen und Codebücher open access zur Verfügung stellt (www.hope.uzh.ch/doca).

Instrumente entwickelt, sodass wenig Standardisierung herrscht, was eine Vergleichbarkeit oder themenübergreifende Evaluation der Wissenschaftskommunikation erheblich erschwert.

Literatur

Akin H (2017) Overview of the science of science communication. In: Jamieson KH, Kahan DM, Scheufele DA, Akin H (Hrsg) The Oxford Handbook of the Science of Science Communication. Oxford University Press, New York, S 1–17

Bal TL, Sharik TL (2019) Web content analysis of university forestry and related natural resources landing webpages in the United States in relation to student and faculty diversity. J Forest 117(4):379–397. https://doi.org/10.1093/jofore/fvz024

Barr S (2011) Climate forums: virtual discourses on climate change and the sustainable lifestyle. Area 43(1):14–22

Bélanger CH, Bali S, Longden B (2013) How Canadian universities use social media to brand themselves. Tert Educ Manag 20(1):14–29. https://doi.org/10.1080/13583883.2013.852237

Berelson B (1952) Content analysis in communication research. Free Press, Glencoe

Bohr J (2020) Reporting on climate change: A computational analysis of U.S. newspapers and sources of bias, 1997–2017. Global Environmental Change 61:102038. https://doi.org/10.1016/j.gloenvcha.2020.102038

Brechman JM, Lee C-J, Cappella JN (2009) Lost in translation? A comparison of cancer-genetics reporting in the press release and its subsequent coverage in lay press. Sci Commun 30(4):453–474. https://doi.org/10.1177/1075547009332649

Brechman JM, Lee C-J, Cappella JN (2011) Distorting genetic research about cancer: From bench science to press release to published news. J Commun 61(3):496–513. https://doi.org/10.1111/j.1460-2466.2011.01550.x

Brosius H-B, Haas A, Koschel F (2016) Methoden der empirischen Kommunikationsforschung: Eine Einführung. Springer VS, Wiesbaden

Bubela TM, Caulfield TA (2004) Do the print media „hype“ genetic research? A comparison of newspaper stories and peer-reviewed research papers. CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne 170(9):1399–1407. https://doi.org/10.1503/cmaj.1030762

Bubela TM, Nisbet MC, Borchelt R, Brunger F, Critchley C, Einsiedel E, Geller G, Gupta A, Hampel J, Hyde-Lay R, Jandciu EW, Jones SA, Kolopack P, Lane S, Lougheed T, Nerlich B, Ogbogu U, O’Riordan K, Ouellette C, Spear M, Strauss S, Thavaratnam T, Willemse L, Caulfield T (2009) Science communication reconsidered. Nat Biotechnol 27:514–518. https://doi.org/10.1038/nbt0609-514

Burns TW, O’Connor DJ, Stocklmayer SM (2003) Science communication: A contemporary definition. Public Underst Sci 12(2):183–202. https://doi.org/10.1177/09636625030122004

Cacciatore MA, Anderson AA, Choi D-H, Brossard D, Scheufele DA, Liang X, Ladwig PJ, Xenos M, Dudo A (2012) Coverage of emerging technologies: A comparison between print and online media. New Media Soc 14(6):1039–1059. https://doi.org/10.1177/1461444812439061

Capriotti P, Carretón C, Castillo A (2016) Testing the level of interactivity of institutional websites: From museums 1.0 to museums 2.0. International Journal of Information Management 36(1):97–104. https://doi.org/10.1016/j.ijinfomgt.2015.10.003
Castillo-Esparcia A, Almansa-Martínez A, Smolak-Lozano E (2015) East European think tanks in social media – towards the model of evaluation of effective communication/PR strategies: Case study analysis. Catalan Journal of Communication & Cultural Studies 7(2):231–250. https://doi.org/10.1386/cjcs.7.2.231_1
Cho M, Schweickart T, Haase A (2014) Public engagement with nonprofit organizations on Facebook. Public Relations Review 40(3):565–567. https://doi.org/10.1016/j.pubrev.2014.01.008
Dudo A, Dunwoody S, Scheufele DA (2011) The emergence of nano news: Tracking thematic trends and changes in U.S. newspaper coverage of nanotechnology. Journalism & Mass Communication Quarterly 88(1):55–75. https://doi.org/10.1177/107769901108800104
Erviti MC, Codina M, León B (2020) Pro-science, anti-science and neutral science in online videos on climate change, vaccines and nanotechnology. Media and Communication 8(2):329–338. https://doi.org/10.17645/mac.v8i2.2937
Fähnrich B (2017) Wissenschaftsevents zwischen Popularisierung, Engagement und Partizipation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer Fachmedien Wiesbaden, Wiesbaden, S 165–182
Fähnrich B (2018a) Digging deeper? Muddling through? How environmental activists make sense and use of science — an exploratory study. JCOM 17(3). https://doi.org/10.22323/2.17030208
Fähnrich B (2018b) Einflussreich, aber wenig beachtet?: Eine Meta-Studie zum Stand der deutschsprachigen Forschung über strategische Kommunikation von Wissenschaftsorganisationen. Publizistik 63:407–426. https://doi.org/10.1007/s11616-018-0435-z
Friedrichsmeier A, Laukötter E, Marcinkowski F (2015) Hochschul-PR als Restgröße. Wie Hochschulen in die Medien kommen und was ihre Pressestellen dazu beitragen. In: Bonfadelli H, Schäfer MS, Kristiansen S (Hrsg) Wissenschaftskommunikation im Wandel. von Halem, Köln, S 128–152
Früh W (2017) Inhaltsanalyse: Theorie und Praxis. UVK Verlagsgesellschaft mbH, Konstanz
Gerhards J, Schäfer MS (2011) Normative Modelle wissenschaftlicher Öffentlichkeit: Theoretische Systematisierung und Illustration am Fall der Humangenomforschung. In: Ruhrmann G, Milde J, Zillich AF (Hrsg) Molekulare Medizin und Medien. Zur Darstellung und Wirkung eines kontroversen Wissenschaftsthemas. VS Verlag für Sozialwissenschaften, Wiesbaden, S 19–40
Guenther L, Bischoff J, Löwe A, Marzinkowski H, Voigt M (2019) Scientific evidence and science journalism. Journal Stud 20(1):40–59. https://doi.org/10.1080/1461670X.2017.1353432
Hara N, Abbazio J, Perkins K (2019) An emerging form of public engagement with science: Ask Me Anything (AMA) sessions on Reddit r/science. PLoS ONE 14(5):e0216789. https://doi.org/10.1371/journal.pone.0216789

Heyl A, Joubert M, Guenther L (2020) Churnalism and hype in science communication: Comparing university press releases and journalistic articles in South Africa. Communicatio 46(2):126–145. https://doi.org/10.1080/02500167.2020.1789184
Jahng MR, Lee N (2018) When scientists tweet for social changes: Dialogic communication and collective mobilization strategies by flint water study scientists on Twitter. Sci Commun 40(1):89–108. https://doi.org/10.1177/1075547017751948
Jarreau PB, Dahmen NS, Jones E (2019) Instagram and the science museum: A missed opportunity for public engagement. JCOM 18(2):1–22. https://doi.org/10.22323/2.18020206
Jensen B (2013) Instagram as cultural heritage: User participation, historical documentation, and curating in Museums and archives through social media 2013 Digital Heritage International Congress (DigitalHeritage). IEEE, S 311–314
Jun J (2011) How climate change organizations utilize websites for public relations. Public Relations Review 37(3):245–249. https://doi.org/10.1016/j.pubrev.2011.04.001
Jünger J, Fähnrich B (2019) Does really no one care? Analyzing the public engagement of communication scientists on Twitter. New Media Soc 22(3):387–408. https://doi.org/10.1177/1461444819863413
Kessler SH (2016) Das ist doch evident!: Eine Analyse dargestellter Evidenzframes und deren Wirkung am Beispiel von TV-Wissenschaftsbeiträgen. Nomos Verlagsgesellschaft mbH & Co, KG, Baden-Baden
Kessler SH, Schäfer MS (2022) Content Analyses in Science Communication Research. In: Oehmer F, Kessler SH, Humprecht E, Sommer K, Castro Herrero L (Hrsg) Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Springer VS, Wiesbaden
Kessler SH, Fähnrich B, Schäfer MS (2020) Science communication research in the German-speaking countries: A content analysis of conference abstracts. Studies in Communication Sciences 19(2):243–251. https://doi.org/10.24434/J.SCOMS.2019.02.012
Kessler SH, Sommer K, Humprecht E, Oehmer F (2022) Manuelle standardisierte Inhaltsanalyse. In: Oehmer F, Kessler SH, Humprecht E, Sommer K, Castro Herrero L (Hrsg) Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Springer VS, Wiesbaden
Knudsen S (2005) Communicating novel and conventional scientific metaphors: a study of the development of the metaphor of genetic code. Public Underst Sci 14(4):373–392. https://doi.org/10.1177/0963662505056613
Kohring M (1997) Die Funktion des Wissenschaftsjournalismus: Ein systemtheoretischer Entwurf. Westdeutscher Verlag, Opladen
Kouper I (2010) Science blogs and public engagement with science: Practices, challenges, and opportunities. JCOM 9(1). https://doi.org/10.22323/2.09010202
Lee NM, VanDyke MS (2015) Set it and forget it: The one-way use of social media by government agencies communicating science. Sci Commun 37(4):533–541. https://doi.org/10.1177/1075547015588600
Lee NM, VanDyke MS, Cummins RG (2017) A missed opportunity?: NOAA's use of social media to communicate climate science. Environ Commun 12(2):274–283. https://doi.org/10.1080/17524032.2016.1269825

Linvill DL, McGee SE, Hicks LK (2012) Colleges' and universities' use of Twitter: A content analysis. Public Relations Review 38(4):636–638. https://doi.org/10.1016/j.pubrev.2012.05.010

Linvill DL, Rowlett JT, Kolind MM (2015) Academic Pinstitution: Higher education's use of Pinterest for relationship marketing. Journal of Relationship Marketing 14(4):287–300. https://doi.org/10.1080/15332667.2015.1093581

Lopera E, Moreno C (2014) The uncertainties of climate change in Spanish daily newspapers: content analysis of press coverage from 2000 to 2010. JCOM 13(1):A02. https://doi.org/10.22323/2.13010202

Lörcher I, Taddicken M (2017) Discussing climate change online. Topics and perceptions in online climate change communication in different online public arenas. JCOM 16(2):A03. https://doi.org/10.22323/2.16020203

Mahrt M, Puschmann C (2014) Science blogging: An exploratory study of motives, styles, and audience reactions. JCOM 13(3). https://doi.org/10.22323/2.13030205

Mellor F (2010) Negotiating uncertainty: asteroids, risk and the media. Public Underst Sci 19(1):16–33. https://doi.org/10.1177/0963662507087307

Metag J, Schäfer MS (2017) Hochschulen zwischen Social Media-Spezialisten und Online-Verweigerern. Eine Analyse der Online-Kommunikation promotionsberechtigter Hochschulen in Deutschland, Österreich und der Schweiz. SCM Studies in Communication and Media 6(2):160–195. https://doi.org/10.5771/2192-4007-2017-2-160

Metag J, Schäfer MS (2019) Hochschulkommunikation in Online-Medien und Social Media. In: Fähnrich B, Metag J, Post S, Schäfer MS (Hrsg) Forschungsfeld Hochschulkommunikation. Springer Fachmedien Wiesbaden, Wiesbaden, S 363–391

Nisbet MC, Brossard D, Kroepsch A (2003) Framing Science: The stem cell controversy in an age of press/politics. The International Journal of Press/Politics 8(2):36–70. https://doi.org/10.1177/1081180X02251047

Peters HP (1994) Wissenschaftliche Experten in der öffentlichen Kommunikation über Technik, Umwelt und Risiken. Kölner Zeitschrift für Soziologie und Sozialpsychologie, S 162–190

Peters HP, Allgaier J, Dunwoody S, Lo Y-Y, Brossard D, Jung A (2013) Medialisierung der Neurowissenschaften. In: Grande E, Jansen D, Jarren O, Rip A, Schimank U, Weingart P (Hrsg) Neue Governance der Wissenschaft. transcript Verlag, S 311–336

Pfenning U (2012) Zur Evaluation von Modellprojekten zur Wissenschaftskommunikation. In: Dernbach B, Kleinert C, Münder H (Hrsg) Handbuch Wissenschaftskommunikation. VS Verlag für Sozialwissenschaften, Wiesbaden, S 341–352

Raupp J, Vogelgesang J (2009) Medienresonanzanalyse: Eine Einführung in Theorie und Praxis. VS Verlag, Wiesbaden

Rödder S, Schäfer MS (2010) Repercussion and resistance. An empirical study on the interrelation between science and mass media. Communications 35(3):249–267. https://doi.org/10.1515/comm.2010.014

Rössler P (2017) Inhaltsanalyse. UVK Verlagsgesellschaft mbH, Konstanz, München

Ruhrmann G, Guenther L, Kessler SH, Milde J (2015) Frames of scientific evidence: How journalists represent the (un)certainty of molecular medicine in science television programs. Public Underst Sci 24(6):681–696. https://doi.org/10.1177/0963662513510643

Schäfer MS (2009) From Public Understanding to Public Engagement: An Empirical Assessment of Changes in Science Coverage. Sci Commun 30(4):475–505. https://doi.org/10.1177/1075547008326943

Schäfer MS (2012) Taking stock: A meta-analysis of studies on the media's coverage of science. Public Underst Sci 21(6):650–663. https://doi.org/10.1177/0963662510387559

Schäfer MS, Fähnrich B (2020) Communicating science in organizational contexts: toward an "organizational turn" in science communication research. JCOM 24(3):137–154. https://doi.org/10.1108/JCOM-04-2020-0034

Schäfer MS, O'Neill S (2017) Frame Analysis in Climate Change Communication: Approaches for Assessing Journalists' Minds, Online Communication and Media Portraya. In: Nisbet MC, Ho SS, Markowitz E, O'Neill S, Schäfer MS, Thaker J (Hrsg) Oxford Research Encyclopedia of Climate Science, Oxford University Press

Schäfer MS, Kristiansen S, Bonfadelli H (2015) Wissenschaftskommunikation im Wandel: Relevanz, Entwicklung und Herausforderungen des Forschungsfeldes. In: Bonfadelli H, Schäfer MS, Kristiansen S (Hrsg) Wissenschaftskommunikation im Wandel. Herbert von Halem Verlag, Köln, S 10–42

Schäfer MS, Kessler SH, Fähnrich B (2019) Analyzing science communication through the lens of communication science: Reviewing the empirical evidence. In: Leßmöllmann A, Dascal M, Gloning T (Hrsg) Science Communication. De Gruyter, Berlin, Boston, S 77–104

Scheufele B, Engelmann I (2009) Empirische Kommunikationsforschung. UVK-Verlagsgesellschaft, Konstanz

Schmid-Petri H, Bürger M (2019) Modeling science communication: from linear to more complex models. In: Leßmöllmann A, Dascal M, Gloning T (Hrsg) Science Communication. De Gruyer, Berlin, Boston, S 105–122

Shea NA (2015) Examining the nexus of science communication and science education: A content analysis of genetics news articles. J Res Sci Teach 52(3):397–409. https://doi.org/10.1002/tea.21193

Shema H, Bar-Ilan J, Thelwall M (2012) Research blogs and the discussion of scholarly information. PLoS ONE 7(5):e35869. https://doi.org/10.1371/journal.pone.0035869

Stocking SH, Holstein LW (2009) Manufacturing doubt: journalists' roles and the construction of ignorance in a scientific controversy. Public Underst Sci 18(1):23–42. https://doi.org/10.1177/0963662507079373

Su LY-F, Scheufele DA, Bell L, Brossard D, Xenos MA (2017) Information-sharing and community-building: Exploring the use of Twitter in science public relations. Sci Commun 39(5):569–597. https://doi.org/10.1177/1075547017734226

Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Venetis CA, Davies A, Ogden J, Whelan L, Hughes B, Dalton B, Boy F, Chambers CD (2014) The association between exaggeration in health related science news and academic press releases: Retrospective observational study. BMJ (Clinical research ed.) 349:g7015. https://doi.org/10.1136/bmj.g7015

Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Bott L, Adams R, Venetis CA, Whelan L, Hughes B, Chambers CD (2016) Exaggerations and caveats in press releases and health-related science news. PLoS ONE 11(12):e0168217. https://doi.org/10.1371/journal.pone.0168217

Taddicken M, Wolff L, Wicke N, Götjen D (2019) Beteiligung und Themenkonstruktion zum Klimawandel auf Twitter. In: Neverla I, Taddicken M, Lörcher I, Hoppe I (Hrsg) Klimawandel im Kopf. Springer Fachmedien Wiesbaden, Wiesbaden, S 229–262

Taddicken M, Reif A, Brandhorst J, Schuster J, Diestelhorst M, Hauk L (2020) Wirtschaftlicher Nutzen statt gesellschaftlicher Debatte? Eine quantitative Framing- Analyse der Medienberichterstattung zum autonomen Fahren. M&K 68(4):406–427. https://doi.org/10.5771/1615-634X-2020-4-406

Taylor M, Kent ML, White WJ (2001) How activist organizations are using the Internet to build relationships. Public Relations Review 27(3):263–284. https://doi.org/10.1016/S0363-8111(01)00086-8

Vestergård GL, Nielsen KH (2016) Science news in a closed and an open media market: A comparative content analysis of print and online science news in Denmark and the United Kingdom. Eur J Commun 31(6):661–677. https://doi.org/10.1177/0267323116674110

Vestergård GL, Nielsen KH (2017) From the preserves of the educated elite to virtually everywhere: A content analysis of Danish science news in 1999 and 2012. Public Underst Sci 26(2):220–234. https://doi.org/10.1177/0963662515603272

Walter S, Lörcher I, Brüggemann M (2019) Scientific networks on Twitter: Analyzing scientists' interactions in the climate change debate. Public Underst Sci 28(6):696–712. https://doi.org/10.1177/0963662519844131

Waters RD, Jamal JY (2011) Tweet, tweet, tweet: A content analysis of nonprofit organizations' Twitter updates. Public Relations Review 37(3):321–324. https://doi.org/10.1016/j.pubrev.2011.03.002

Waters RD, Burnett E, Lamm A, Lucas J (2009) Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. Public Relations Review 35(2):102–106. https://doi.org/10.1016/j.pubrev.2009.01.006

Weingart P (2012) The lure of the mass media and its repercussions on science. Theoretical considerations on the „Medialization of Science". In: Rödder S, Franzen M, Weingart P (Hrsg) The Sciences' Media Connection – Public Communication and its Repercussions. Springer Netherlands, Dordrecht, S 17–32

Wicke N (2021) Citizen Science – eine ›erfolgreiche‹ Entwicklung in der Wissenschaftskommunikation? In: Milde J, Welzenbach-Vogel IC, Dern M (Hrsg) Intention und Rezeption von Wissenschaftskommunikation. Herbert von Halem Verlag, Köln, S 177–206

Wicke N (2022) Content Analysis in the Research Field of Science Communication. In: Oehmer F, Kessler SH, Humprecht E, Sommer K, Castro Herrero L (Hrsg) Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Springer VS, Wiesbaden

Winters M, Larsson A, Kowalski J, Sundberg CJ (2019) The association between quality measures of medical university press releases and their corresponding news stories-Important information missing. PLoS ONE 14(6):e0217295. https://doi.org/10.1371/journal.pone.0217295

Wirth W, Sommer K, Wettstein M, Matthes J (Hrsg) (2015) Qualitätskriterien in der Inhaltsanalyse. Halem, Köln

Yang A, Taylor M (2010) Relationship-building by Chinese ENGOs' websites: Education, not activation. Public Relations Review 36(4):342–351. https://doi.org/10.1016/j.pubrev.2010.07.001

Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, Ravaud P (2012) Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. PLoS Med 9(9):e1001308. https://doi.org/10.1371/journal.pmed.1001308
Yeon HM, Choi Y, Kiousis S (2007) Interactive communication features on nonprofit organizations' webpages for the practice of excellence in public relations. Journal of Website Promotion 1(4):61–83. https://doi.org/10.1300/J238v01n04_06
Zerfaß A, Volk SC (2019) Evaluationstools. In: Zerfaß A, Volk SC (Hrsg) Toolbox Kommunikationsmanagement. Springer Fachmedien Wiesbaden, Wiesbaden, S 181–218
Zhang Y, O'Halloran KL (2013) 'Toward a global knowledge enterprise': University websites as portals to the ongoing marketization of higher education. Crit Discourse Stud 10(4):468–485. https://doi.org/10.1080/17405904.2013.813777
Ziegler R, Hedder IR, Fischer L (2021) Evaluation of science communication: Current practices, challenges, and future implications. Front Commun 6. https://doi.org/10.3389/fcomm.2021.669744

Dr. Sabrina H. Kessler ist Senior Research and Teaching Associate in der Abteilung für Wissenschaftskommunikation des Instituts für Kommunikationswissenschaft und Medienforschung der Universität Zürich und Mitglied der Jungen Akademie Schweiz. Sie promovierte in der Kommunikationswissenschaft an der Friedrich-Schiller-Universität Jena. Ihre Forschungsschwerpunkte umfassen Wissenschafts- und Gesundheitskommunikation sowie Online-Recherche-, Selektions- und Rezeptionsverhalten.

Nina Wicke war wissenschaftliche Mitarbeiterin am Institut für Kommunikationswissenschaft der Technischen Universität Braunschweig. Sie hat zu Qualitätskriterien von Wissenschaftskommunikation aus Publikumssicht promoviert. Ihre Forschungsinteressen umfassen Evaluation von Wissenschaftskommunikation, Klimawandel- und Expert:innenkommunikation sowie Citizen Science.



Praxisbeitrag: Anwendungsbeispiel zur Integration inhaltsanalytischer Betrachtungen in Multi-Methoden-Forschungsstrategien im Bereich der Wissenschaftskommunikation

Rüdiger Goldschmidt und Oliver Scheel

Zusammenfassung

Der Beitrag skizziert die Potenziale inhaltsanalytischer Betrachtungen als Bestandteil der Begleitforschung von Initiativen im Bereich Wissenschaftskommunikation und speziell in Bezug auf den Einsatz interaktiver Online-Designs. Inhaltsanalytische Betrachtungen erwiesen sich als zielführende Ergänzung des Multi-Method Designs der Forschungsstudie, wobei die Verzahnung aller Methoden eine Rolle spielt. Die vorgestellte explorative Teilstudie erarbeitet auch inhaltliche Erkenntnisse. Online-Verfahren wie Foren oder Diskussionstische zeigten viel Potenzial für die Wissenschaftskommunikation insbesondere dann, wenn die Auswahl des Designs unter Berücksichtigung auf die Zielsetzung des Dialogverfahrens erfolgt. Mit den Ergebnissen deutet sich an, dass die Frage welche Kommunikationsmodi (deliberativ-konstruktiv, aggregierend oder additiv) in einem Dialogdesign entwickelt werden sollen bzw. können wichtiger ist als die Frage der Umsetzung als Online versus Präsenzveranstaltung.

R. Goldschmidt (✉)
Dialogik gemeinnützige Gesellschaft für Kommunikations- und Kooperationsforschung mbH, Stuttgart, Deutschland
E-Mail: goldschmidt@dialogik-expert.de

O. Scheel
ZIRIUS Universität Stuttgart, Stuttgart, Deutschland
E-Mail: oliver.scheel@zirius.uni-stuttgart.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_15

1 Zielsetzung und Erläuterungen zur Zielsetzung

Der Beitrag verfolgt das Ziel, Potenziale inhaltsanalytischer Begleitforschung in der Wissenschaftskommunikation am Beispiel von Untersuchungen einer konzeptuellen wissenschaftlichen Fragestellung zu skizzieren. Viele sehen dabei die „methodische Mechanik" im Vordergrund, d. h. die einzelnen Methoden wie Befragung oder Interview. Dieses Handwerkszeug ist unverzichtbar. Wichtig ist zudem, Forschung und Forschungsmethoden im jeweiligen Kontext und Anwendungsgebiet zu betrachten. Der Beitrag greift einige Ergebnisse aus einer umfangreicheren Studie heraus, um zu zeigen, wie inhaltsanalytische Betrachtungen in Forschungsvorhaben und ggf. in Verbünde von Methoden, sogenannte Multi-Methoden-Ansätze, integriert werden können. Aus Platzgründen widmet sich der Beitrag dabei zentralen Aspekten der Gesamtstudie und stellt nur begrenzt Details vor. Die einzelnen Kapitel und Unterabschnitte sind nach einem verbreiteten „Standard" für empirische Forschungsstudien gegliedert, um den Aufbau der Forschung für viele Leser:innengruppen aufzubereiten: Forschende kennen normalerweise diesen Standard. Diejenigen, die sich mit Forschung vertraut machen wollen, gewinnen einen Eindruck von den Funktionalitäten und Abläufen. Forschungsaktivitäten gehen von wissenschaftlichen und praktischen Fragestellungen aus, was im Abschnitt zur Problemstellung umrissen wird (Abschn. 2.1). Zur „konzeptuellen Basis" gehören außerdem die begrifflichen Ansatzpunkte sowie Annahmen (Abschn. 2.2). Nur dieser Bezug zu einer konzeptuellen Basis macht empirische Forschungsstudien und Methoden überhaupt sinnhaft interpretierbar. Nach Erläuterungen zu den eingesetzten Forschungsmethoden (Kap. 3) werden einige Ergebnisse vorgestellt (Kap. 4). Darauf folgen Reflexion, Bewertung und Interpretation der Ergebnisse sowie der eingesetzten Methoden. Ein kurzer Ausblick fasst interessante Forschungsfragen oder Problemstellungen für zukünftige Forschung zusammen. Die konkrete Ausgestaltung von Forschungsaktivitäten läuft jedoch nicht mechanisch nach simplen Schemata ab. Forschung ist dynamisch und individuell. Forschungsstudien unterscheiden sich daher in ihrer konkreten Umsetzung. Das hier erläuterte Schema lässt sich grundlegend auf viele Forschungssettings anwenden.

Die vorliegende explorative Studie ergänzt die Befunde einer umfassenderen Forschungsstrategie. Der sogenannte Formatvergleich zielte darauf ab, im Feld

der Wissenschaftskommunikation sieben Dialog- und Beteiligungsformate[1] systematisch und auf empirischer Basis miteinander zu vergleichen. Fragestellungen waren z. B., ob die Teilnahme an den Formaten das Interesse der Teilnehmenden an Wissenschaft und wissenschaftlichen Fragestellungen stärkt. Untersucht wurde zudem, ob die Teilnahme die Aufgeschlossenheit gegenüber der Wissenschaft beeinflusst sowie Sachwissen und Urteilsfähigkeit in Bezug auf das diskutierte Thema steigert. Daneben gab es eine Anzahl von weiteren Forschungsfragen, die aus Platzgründen nicht weiter ausgeführt werden.

Die empirischen Untersuchungen stützen sich auf einen umfangreichen Verbund von sich ergänzenden Forschungsmethoden, einen sogenannten Multi-Methoden-Ansatz (vgl. Abschn. 3 in diesem Beitrag). Ein Verbund aus verschiedenen Organisationen im Bereich der Wissenschaftskommunikation, darunter Wissenschaft im Dialog sowie ZIRIUS von der Universität Stuttgart (damals ZIRN), teilten sich die Aufgaben der Organisation der Veranstaltungen sowie der Begleitforschung. Pro Format wurde eine bestimmte Anzahl von Veranstaltungen durchgeführt: 4 Schüler:innenparlamente, 10 Schüler:innenforen, 10 Junior Science Cafés, 4 Bürger:innenkonferenzen, 1 Konsensuskonferenz, 2 Bürger:innenausstellungen sowie diverse Onlineaktivitäten. Die unterschiedliche Anzahl von Veranstaltungen bei den Formaten erklärt sich aus den teilweise verschiedenen, bei manchen Veranstaltungen hohen Aufwendungen. Die Konsensuskonferenz war vergleichsweise kostenintensiv, sodass sie nur einmal durchgeführt werden konnte. Die meisten Veranstaltungen basierten auf einem Grunddesign, das den Teilnehmenden erlaubte, sich persönlich von Angesicht zu Angesicht miteinander auszutauschen. Es handelte sich um sogenannte Präsenzformate. Die im Vorhaben durchgeführten Online-Designs waren damals noch relativ neu. Die Forschung führte zu relevanten Befunden insbesondere in Bezug auf die Möglichkeiten sich miteinander auszutauschen, d. h. verschiedene Modi der Kommunikation. Insofern fokussiert der vorliegende Beitrag auf die vergleichende Forschung zu diesen Online-Designs.

[1] „Formate“ sind eher abstrakte allgemeine Prozessformen, die sich aus empirischen Fällen ableiten lassen. So schließt ein Format jeweils Dialog- und Beteiligungsverfahren mit gleichen oder ähnlichen Merkmalsausprägungen ein. Das Format der „Konsensuskonferenz” lässt sich bspw. von der „Bürgerkonferenz” unterscheiden in Bezug auf die Anzahl von Teilnehmenden sowie anhand von Merkmalen der Entscheidungsprozesse. „Designs” repräsentieren Möglichkeiten, wie die Formen von Interaktionen und Kommunikation in Formaten organisiert sind. Eine Bürgerkonferenz kann z. B. als Präsenz- oder Online- oder Hybrid-Veranstaltung umgesetzt werden.

2 Konzeptuelle Basis

2.1 Problemstellung

Das Aufkommen neuer Technologien eröffnete für Forschung und Anwendung in der Wissenschaftskommunikation nicht nur neue Möglichkeiten, sondern auch Fragen. Auf die mit der Entwicklung des Internets verbundenen tiefgreifenden sozialen Veränderungsprozesse und -potenziale wurde seit langem hingewiesen (Baek et al. 2011, Carlitz et al. 2005; Galston 1999; Turkle 1995). Diese Dynamik beeinflusst ebenfalls das Feld der Partizipation und Deliberation (Wright et al. 2007). Argumente für oder gegen Virtualisierung von Dialogen liegen schon länger vor. Einerseits bringen Autoren die niedrigen Kosten für Online-Formate bzw. Foren an (Carlitz et al. 2005; Quinlan et al. 2015; Wright et al. 2007). Zum Beispiel erübrigen sich Anreise und Unterbringung von Teilnehmenden. Jedoch fallen andere Kosten an, bspw. für die Software und Systemadministration usw. Die Interaktionsqualität wird ebenfalls thematisiert. Teilnehmende an Online-Diskussionstischen sind eben nicht mehr örtlich in Präsenzveranstaltungen gebunden (Zittel 2001; Vor- und Nachteile zusammenfassend Quinlan et al. 2015). Ein zeitlich ungebundener Austausch z. B. in Online-Foren könnte sogar reflexiver und rationaler sein als Tischdiskussionen. Über Anonymität ließe sich womöglich der freie Austausch fördern. Dagegen könnte die Anonymität ebenso das „Lurking" intensivieren (Amichai-Hamburger 2016), den unhöflichen Umgang miteinander oder stereotype Argumentationen (Quinlan et al. 2015). Recht früh wurde in der Diskussion auch auf das Potenzial zur Polarisierung von Perspektiven und Diskursen hingewiesen (zusammenfassend Wright et al. 2007), z. B. wenn Personen nur in Foren teilnehmen, in denen ihre Perspektive dominiert. So könnten Online-Diskussionen die offene Deliberation geradezu aushöhlen (Wright et al. 2007). Viele Fragen in Bezug auf den technisch vermittelten Austausch sind also noch offen (Barber 1995; Fuchs 2007; Rosenfield 1998; Ravetz 1998). Die verfügbaren Befunde gerade zur Online-Partizipation sind ambivalent (Quinlan et al. 2015). So besteht hoher Bedarf an empirischer Forschung (Carpini et al. 2004). Es braucht differenziertere und integrativere Forschungsstrategien, welche auf die verschiedenen Formen von Partizipation eingehen (Wright et al. 2007). Inhaltsanalysen bilden einen wichtigen Baustein in Bezug auf die Forschungsmethoden.

2.2 Konzeptueller Ansatz und Annahmen

Die zentrale Annahme der Forschungsstrategie des Formatvergleiches lautete, dass Unterschiede zwischen Formaten hinsichtlich der prozeduralen Merkmale bestehen, z. B. hinsichtlich der Kommunikations- und Interaktionsmuster. Dies geht mit Unterschieden zwischen Formaten bezüglich der Wirkungspotenziale auf die Teilnehmenden einher. Mit den vorliegenden Ergebnissen anderer Studien deutet sich bereits an, dass es einen Erfolgsfaktor darstellt, Unterschiede zwischen Designs zu berücksichtigen (für Online Foren Jensen 2003; Quinlan et al. 2015; Wright et al. 2007). Die Forschung zu technisch vermittelten Lernsituationen stellte heraus, dass die Qualität von Interaktionen grundsätzlichen Einfluss auf das Lernen besitzt, wobei Interaktionsmuster gesucht wurden, die Gruppenarbeit fördern (Oliveira et al. 2011). Die Grundannahme des Formatvergleiches wurde für die untersuchten sieben Dialog- und Beteiligungsformate schon an anderer Stelle bestätigt (Goldschmidt 2018; Goldschmidt et al. 2012). Die Unterschiede zwischen Online- sowie Präsenzdesigns lassen sich mit Fokus auf die Interaktions- bzw. Kommunikationsmodi noch detaillierter herausarbeiten.

Konzeptuell lässt sich die Unterscheidung von Formaten und Designs auf die Differenzierung von aggregierenden versus deliberativen Kommunikations- und Entscheidungsverfahren zurückführen (Cohen 1999a, 1999b; Elster 1999; Feindt 2001; Gerhards 1998; Gerhards et al. 1998; Habermas 1998, Papadopoulos et al. 2007; Wright et al. 2007)[2]. Empirisch scheinen beide Modi in Online-Foren oder Chats aufzutreten (Stromer-Galley et al. 2015; Wolfgang et al. 2015). Deliberation steht für den offenen, intensiven und auf das Gegenüber orientierten, stark reflexiven Verständigungsprozess. Die aufeinander bezogenen Argumente werden in einer Art Schlagabtausch gegenseitig bewertet und versucht, z. B. durch Überzeugen des Gegenübers, die Einigung herbeizuführen. Der aggregierende Modus bildet schlicht das Gegenteil des deliberativen Modus. Auch wenn einzelne Individuen ihre Haltung ändern können, bleiben die Meinungen über mehrere Teilnehmende ohne gegenseitig bezogene gemeinsame, verständigungsorientierte Reflexion in Form von Meinungspolen relativ unverbunden bestehen. Im Ergebnis kann so keine gemeinsame Position entwickelt

[2] Deliberation wurde konzeptuell eingegrenzt (Goldschmidt 2014; Bächtiger et al. 2018). Trotzdem besteht Bedarf für die weitere Differenzierung des Konzeptes z. B. in Bezug auf die Frage, welche konkreten Typen bzw. Formen des Austausches existieren könnten (Bobbio 2010).

werden, sondern allenfalls eben ein Aggregat der Meinungspole. Abstimmungen oder teilweise Verhandlungen sind Beispiele für solche Prozesse.

Die Diskussion um die Typisierung von Gesprächsmodi in Online- und Präsenz-Designs besitzt gerade für die Wissenschaftskommunikation hohe Relevanz und zwar in konzeptueller, methodischer und praktischer Hinsicht. Viele erleben derzeit konstruktive Verständigungsprozesse z. B. in Online-Konferenzen. Vor dem Hintergrund eskalierender Diskussionen in Foren oder Twitter-Chats aktualisieren sich jedoch auch die in einigen früheren Beiträgen geäußerten Zweifel, ob die Online-Kommunikation einen wirklich deliberativen Austausch zulässt (Fuchs 2007; Galston 1999; Quinlan et al. 2015; Ravetz 1998; Rosenfield 1998; Wilhelm 2000). Viele der Kontrastierungen folgen dabei – oft implizit – der Unterscheidung entlang der Linie, dass sich die Dialogqualität primär daran entscheidet, ob ein Austausch technisch vermittelt ist oder nicht. Eine solche Ausgangsthese, grade in so einer generellen und absoluten Form, muss eher zurückgewiesen werden. Technologie *für sich allein*, wie es der technologische Determinismus behauptet, definiert nichts (Wright et al. 2007). Trotzdem bleibt die Aufgabe, empirisch differenzierter zu untersuchen, ob sich tatsächlich Designs z. B. in Bezug auf die Kommunikationsmuster unterscheiden und welche Rolle dabei die technische Vermittlung als Faktor spielt. Diese Frage ist grundlegender Natur und fordert mehr als nur eine Forschungsstudie. Aber eine explorative Studie kann hier Richtungen und Tendenzen aufzeigen.

Deliberative bzw. partizipative Veranstaltungsdesigns lassen sich über zahlreiche Dimensionen vergleichen (Goldschmidt 2018). In Anpassung an die inhaltlichen und methodischen Fragestellungen sowie die Fokussierung auf die Online-Designs wurde für diesen Beitrag der Aspekt der *Intensität und Interaktivität des Informationsaustausches* gewählt (Quinlan et al. 2015; Wright et al. 2007; Abschn. 4 in diesem Beitrag). Hier können die Länge der Interaktionen bzw. der Argumentationsketten zwischen den Beitragenden sowie die Verteilung der Teilnehmenden und ihrer Handlungen, z. B. das Verhältnis von Lese- und Schreibaktivitäten in einem Forum, untersucht werden. Berücksichtigt werden außerdem die Inhalte. Die Kommunikationsmodi stellen letztlich Grundformen dar, wie Teilnehmende Informationen und Perspektiven austauschen. Die Exploration der Muster über die genannte Dimension ermöglicht zumindest langfristig Aussagen bzw. Studien zu Wirkungspotenzialen dieser Muster, z. B. zur Frage, ob der Austausch die Informiertheit von Teilnehmenden oder deren Urteilsfähigkeit stärkt.

3 Vorgehen und Einbettung der inhaltsanalytischen Forschung

Die hier vorgestellte explorative Studie stellt zwei Typen von Online-Foren[3], ein Online-Table-Design[4] sowie Tisch-Diskussionen „konventioneller" Präsenzveranstaltungen einander gegenüber. Mit Verweis auf die Zielsetzung und die umfassendere Hauptstudie konzentriert sich dieser Beitrag stärker auf die Online-Designs und hier auf die Foren. Zur Zeit der Studie 2012 wurden erste Programme zur Durchführung von Online-Diskussionen angeboten, die aktueller Konferenzsoftware ähnelten. Die Designs wurden erprobt und vergleichend untersucht. Diese Forschung stand damals noch relativ am Anfang, wodurch die Studie ihren explorativen Charakter erhielt.

Der empirisch-systematische Formatvergleich (Goldschmidt 2014, 2018; Goldschmidt et al. 2012) ist eine Forschungsstrategie, die Dialog- und Beteiligungsformate mittels eines Multi-Method Designs (Denzin 2009; vgl. auch Gabriel et al. in diesem Band) über mehrere konzeptuelle Dimensionen hinweg miteinander vergleicht. Der Schwerpunkt der bisherigen Studien lag auf dem Vergleich von Präsenzformaten basierend auf mehrwelligen Befragungen der Teilnehmenden, Interviews mit einbezogenen Sachverständigen und Moderierenden sowie nicht-teilnehmenden Beobachtungen. Ergänzend führte das Forschungsteam Medienanalysen durch, z. B. von verfügbarer Berichterstattung über das Projekt bzw. die Formate in Radio oder Zeitungen (vgl. Tab. 1). Das Gesamtdesign der Studie war vergleichsweise komplex. Diese Komplexität erhöhte sich noch dadurch, dass die Datenerhebung in jeder Veranstaltung häufig die Perspektiven verschiedener Zielgruppen erfasste. Allgemein bilden Multi-Method-Designs einen zentralen Entwicklungspfad moderner sozialwissenschaftlicher Analyse. Die

[3] Online-Foren, sogenannte Message- oder Diskussion-Boards, sind über Webseiten vermittelte Kommunikationsmedien bei denen sich die Teilnehmenden durch das Veröffentlichen kurzer Texte (Posts) miteinander austauschen. Die Posts sind fast immer öffentlich einsehbar, um Leser:innen zu ermuntern, sich selbst einzubringen. Die Texte fallen üblicherweise länger aus als bei Social-Media-Plattformen wie Twitter oder Facebook (siehe auch https://www.pcmag.com/encyclopedia/term/internet-forum). Das sogenannte begleitende Forum bildete zusammen mit zwei parallellaufenden Bürger:innenkonferenzen eine Prozesskette. Das alleinstehende Forum war mit keinem anderen Format verbunden.

[4] Online-Table-Designs sind klassische Diskussionsrunden, allerdings ohne örtliche Präsenz der Teilnehmenden. Sie nutzen digitale Medien. Standard sind Audio- und Video-Liveübertragung sowie Chat mit Emoji-Übertragung.

Tab. 1 Übersicht der Forschungsmethoden des Format- und Designvergleiches. (Quelle: Goldschmidt 2018, angepasst)

Datenbasis	Instrumente	Im Fokus
Untersuchung von subjektiven Eindrücken	Vollstandardisierte Befragung	Teilnehmende
	Halbstandardisierte Interviews (Gruppen-, Einzelinterviews)	Teilnehmende, Sachverständige, Moderator:innen
Untersuchung von objektivierten Daten	Strukturierte Beobachtung (teil- und vollstrukturiert)	Möglichst alle Veranstaltungsprozesse aller Formate, darunter Online- und Präsenztischdiskussionen
	Medienanalyse	Verfügbare Quellen des Pressespiegels
	Inhaltsanalyse Betrachtung von Online-Nutzungsdaten	Serverdaten aus Online-Foren z. B. Posts, Zugriffszahlen usw.

folgenden Ausführungen schildern, warum inhaltsanalytische Verfahren gerade für die Untersuchung von Online-Designs eingebunden wurden.

Die systematische Forschungsstrategie forderte eine möglichst vollständige Datenerhebung über alle Designs. Jedoch wäre jede Datenerhebung sinnlos, wenn sie nicht zu den individuellen Erhebungssettings also den Dialogveranstaltungen passt. Jede Multi-Method-Strategie ist nur dann erfolgreich, wenn die Forschenden mit diesen Spannungsverhältnissen umgehen und zielführende Kombinationen von Methoden schaffen, sodass die Erkenntnisse verschiedener Methoden sich gegenseitig befruchten. Für den konkreten Fall war die Einbindung inhaltanalytischer Methoden in den Verbund der Forschungsmethoden zielführend. Die Tischdiskussionen in Präsenz- sowie Online-Designs wurden „konventionell" z. B. über eine vorstrukturierte nicht-teilnehmende Beobachtung (Goldschmidt 2016) mit Protokollen und unterstützend mit Videoaufnahmen begleitet. Eine solche Begleitung war bei Online-Designs wie Online-Foren nicht möglich. Die Datenerhebungen bei Online-Designs fanden technisch vermittelt durch die Dokumentation von Informationen auf Servern statt. Diese Nutzungsdaten erlaubten somit die umfassende und präzise Abbildung von Aktivitäten, u. a. über Zeitstempel. Die Daten waren ‘ready-for-analysis’ (Black 2009).

Durch die Bindung an die Nutzungsdaten bzw. Beobachtung richteten sich die Analysen dieser Teilstudie damit stärker auf prozedurale Aspekte und die Interaktionen. Befragungen, um die subjektiven Haltungen der Teilnehmenden eingehender zu untersuchen, waren nur dann möglich, wenn z. B. das Online-Forum begleitend an eine Präsenzveranstaltung angeschlossen war, in welcher Daten über die Hauptstudie gesammelt werden konnten. Die Nutzung des alleinstehenden Forums war teilweise anonym, was die Datenerhebung erschwerte. Wie können nun die auf Servern gesammelten Daten inhaltsanalytisch genutzt werden?

Nutzer:innen hinterlassen mit ihren Aktivitäten in Foren Spuren z. B. durch Posts. Für die Analyse können die Häufigkeiten von Ereignissen wie Posts in den Foren ausgezählt werden, insgesamt oder bspw. nach Zeitabschnitten in Verbindung mit den Zeitstempeln oder pro Thread (Diskussion zu einem Thema im Forum). Eine Analyseeinheit bilden Argumentationsketten, also die Betrachtung der Beiträge zu einem Thema in ihrer zeitlichen Abfolge. Des Weiteren können bei den Beiträgen textstrukturelle Merkmale wie die Textlänge untersucht werden. Das Forschungsteam wertete außerdem die Inhalte von Posts qualitativ aus, z. B. um Kernthemen der Threads und Posts zu typisieren oder um den Anteil von aggressiven Formulierungen zu bestimmen. Insgesamt diente das Datenmaterial zur Beschreibung und Typisierung der Kommunikations- und Interaktionsmuster jedes Designs und letztlich dem Vergleich der Designs. Im Ergebnisteil des Beitrags findet sich eine Auswahl von Befunden für die oben fokussierte Dimension der Intensität und Interaktivität, um Analysemöglichkeiten zu skizzieren. Aus Platzgründen fokussiert der nachfolgende Abschnitt auf die inhaltsanalytischen Betrachtungen der Forendiskussion und stellt andere Ergebnisse aus der Gesamtstudie zusammenfassend dar.

4 Empirische Ergebnisse zur Intensität des Austausches

Der folgende Abschnitt skizziert die empirisch beobachteten Interaktionsmodi der untersuchten Prozessdesigns. Neben den konzeptuell zu erwartenden Modi der aggregierenden versus deliberativen bzw. konstruktiven Kommunikation wurde der „additive“ Modus festgestellt, der bisher kaum diskutiert ist. Die Designs unterschieden sich bzgl. der festgestellten Gewichtung der drei Interaktionsmodi:

- *Online-Foren:* Für das alleinstehende Forum waren intensive interaktive Dialoge eher untypisch. Die 158 untersuchten Threads hatten im Durchschnitt 4,43 Replies. Dabei konzentrierten sich die Aktivitäten in einigen

Threads. 16 % der 158 Threads erhielten überhaupt keine Replies, 35 % wiesen ein bis zwei auf (niedrige Aktivität). Bei 43 % der Threads zeigte sich eine mittlere Aktivität bis zu 13 Replies. Und nur 5 % wiesen eine höhere Intensität über dreizehn Replies auf. Der längste Thread bestand aus 23 Replies. Hierbei wurde noch ein verstärkender Effekt festgestellt: Die meistdiskutierten Threads wurden von den Moderierenden zur sogenannten „Idee des Monats" gekürt. Sachverständige wurden eingeladen, diese Ideen bzw. die Diskussion zu kommentieren. Nicht alle Threads waren erfolgreich, denn das Mittel lag hier bei 13.25 Replies pro Thread, einschließlich der Expert:innenmeinung. Doch insgesamt stellte die Analyse heraus, dass die Expert:innenkommentare ausführlicher ausfielen und weitere Posts von Nutzer:nnen auslösten. Die Möglichkeit, sich mit Expert:innen auszutauschen, beförderte also die Motivation der Nutzer:innen und letztlich die Intensität des Austauschs. Zudem fachten die Moderierenden mit ihren Aktivitäten die Diskussionen an. Über alle Threads des Forums betrachtet zeigte sich die vergleichsweise niedrige Intensität von Foren daran, dass bei den Aktivitäten das Lesen der Threads gegenüber den aktiven Beiträgen durch eigene Posts dominierte. Das traf ebenfalls auf die Ideen des Monats zu. Beim Thread „Energieproduktion in Fitnesszentren" wurde fast zehnmal so häufig gelesen wie gepostet bzw. erwidert (184 zu 19, Rate = 9,68:1). Bei Threads von mindestens mittlerer Intensität fiel das Verhältnis kleiner aus. Bei solchen Threads lagen die Verhältnisse Lesen: Posten im Bereich von 3:1 bis 5:2. Generell wies die Analyse auf eine relativ niedrige Gesamtaktivität im Forum hin. Nach strukturellen Gesichtspunkten betrachtet, stockte also im Forum der Zufluss neuer Informationen in den Austausch. In Bezug auf die Textstruktur waren die meisten Replies kurzgehalten, z. B. „gute Idee". Gelegentlich stoppte der Austausch abrupt. Diese eher kurzatmige Form des Austausches unterstützte die Nutzer:innen kaum dabei, Gedanken gemeinschaftlich weiterzuentwickeln und ein kollektives Verständnis der Themen aufzubauen. Von einem bzw. vom eigenen Meinungskern abweichende Meinungen wurden häufig nicht weiter adressiert und blieben damit einfach stehen.[5] Aufgrund dieses spezifischen Charakters der Kommunikation deutete alles auf einen eigenen Typus des Austauschs hin. Über alle Daten gesehen

[5] Zum Beispiel brachte ein Teilnehmer die Idee darüber ein, die Kernspaltung zur zukünftigen Energieversorgung zu nutzen. Dies erregte Gegenwehr von einem anderen Nutzer, welcher die Idee unsinnig fand und seine Sicht mit Links zu drei YouTube Videos zu belegen versuchte. Dies beendete jegliche Diskussion.

zeichnete sich der „additive“ Kommunikationsmodus durch das Aneinanderketten einzeln für sich stehender, isolierter Beiträge aus. Einige der Replies in einer Argumentationskette bezogen sich nicht einmal auf den ursprünglichen Thread. Die Perspektiven blieben unvermittelt quasi als *„Teile-Haufen“* liegen. Überraschend war die empirische Relevanz des additiven Modus im alleinstehenden Forum mit 52,2 % aller 158 Threads. Jedoch fanden sich auch die beiden konzeptuell herausgestellten Modi, z. B. der aggregierende Modus mit etwas über einem Viertel der Threads (26,4 %). Beim aggregierenden Kommunikationsmodus bezogen sich die Posts einer Argumentationskette auf das Threadthema und häufiger auch aufeinander. Jedoch dienten die Darstellungen dazu, bestimmte Perspektiven zu stärken, was die Reflexion von verbindenden Elementen einschränkte. Es bildeten sich somit *Fraktionen bzw. Meinungsaggregate*. Ein geringer Anteil von 15,7 % der Threads wurden als konstruktiv bzw. deliberativ eingestuft.[6] Über einen intensiven, wechselseitig bezogenen Austausch in einer tatsächlichen Kette von Argumenten innerhalb des Thread-Themas wurden die zu Beginn eingeführten Ideen von den Beitragenden gemeinsam weiterentwickelt. Die Beiträge konstruktiver Argumentationen waren gemeinhin länger und in Einzelheiten ausgearbeiteter als die der anderen Modi.

- Das *„begleitende Forum”* unterschied sich vom alleinstehenden Forum durch noch deutlichere Unterschiede zwischen den Anteilen von Lesen und Posten. Die Verhältnisse von Lesen: Posten lagen zwischen 17:1 beim Thema „Windenergie“ und 40:1 beim Thema „Kernenergie“. Die Threads besaßen dezidiert höhere Reichweite bei eingeschränkter aktiver Beteiligung. Trotz der Gemeinsamkeiten im Grundcharakter beider Foren-Designs gab es interessante Abstufungen. Das begleitende Forum erweiterte und ergänzte die Diskussionen der Präsenzveranstaltung. Es fungierte als Informationsquelle für offene Fragen. Das zeigten direkte inhaltliche Bezüge von Forenbeiträgen auf Diskussionen der Präsenzveranstaltung (36 von 99 Beiträgen) neben der erwähnten hohen Rate von Lesezugriffen. Die Beiträge und Argumentationsketten des begleitenden Forums waren elaborierter bzw. reflexiver als beim alleinstehenden, was sich auch an einem geringeren Anteil des additiven Modus zeigte.
- Nach den Beobachtungsergebnissen ähnelten sich *Online- und Präsenz-Tischdesigns* in Bezug auf viele Merkmale. Der konstruktive Kommunikationsmodus dominierte nach den Ergebnissen der Beobachtung am Referenztisch

[6] Die restlichen 5,7 % der Posts waren nicht eindeutig zuzuordnen.

A der Präsenzveranstaltung. Die Teilnehmenden entwickelten die Dialogergebnisse gemeinsam. Gegensätzliche Meinungen und Argumente wurden aktiv in der Gruppe reflektiert. So bildeten sich im Dialog relativ dichte Ketten von gemeinsam reflektierten Argumenten, auf deren Basis die Ergebnisse abgeleitet wurden.[7] Die Teilnehmenden von Tisch A würdigten Beiträge, z. B. wenn Repräsentant:innen die Tischergebnisse während der Plenarsitzungen der Konferenz vorstellten oder sogar die Rolle von Co-Moderator:innen übernahmen und die Tischdiskussion zusammenfassten. Die konstruktiv verlaufenden Tischdiskussionen waren allerdings nicht gefeit gegen einen Wechsel in andere Kommunikationsmodi wie den aggregierenden Austausch. Beim Referenztisch A gab es eine Konfliktsituation, die sich auf die Diskussion anderer Themen auswirkte. Ein Teilnehmer erinnerte jedoch an die bisher gemeinsam geleistete Arbeit und schlug vor, die gegensätzlichen Perspektiven in die Ergebnisse aufzunehmen (Aggregat). Insgesamt erwies sich die Diskussion an Tisch A mit dem offenen, interaktiven und intensiven Charakter als typischer Fall des konstruktiven bzw. deliberativen Modus‘. Nur an Referenztisch B der Präsenzveranstaltung dominierte der aggregierende bzw. zeitweise sogar additive Kommunikationsmodus. Das beeinflusste die Ergebnisqualität.[8]
- Die *Online-Tisch-Diskussion* unterschied sich hinsichtlich der Intensität kaum und wenn nur in Nuancen von der Kommunikation an den Referenztischen der Präsenzveranstaltung. Die Teilnehmenden nutzten einen „Voice-Channel“ als Hauptmedium des Austausches. So besaß eine Person das Rederecht. Während an Präsenztischen non-verbale Kommunikation mit Gestik und Mimik gegenwärtiger war, bot die Diskussion am Online-Tisch keine Live-Videoporträts wie bei heutiger Konferenzsoftware üblich (Abb. 1). Die

[7] Zum Beispiel brachte ein Teilnehmer sein Wissen über ein individuelles Untergrund Transportsystem für Personen (PRT) ein. Die Gruppe diskutierte die Informationen und nahm sie in die Gruppenergebnisse auf.

[8] Die am Tisch B von den Teilnehmenden eingebrachten Perspektiven wurden selten in der Gruppe reflektiert. Die jeweils Beitragenden verteidigten und wiederholten ihre Position ohne intensiver auf andere einzugehen. Damit blieb der Tischrunde nur, sich bei der Ergebnisentwicklung auf den jeweils kleinsten gemeinsamen Nenner von teilweise isolierten Perspektiven zu einigen. Im Gegenteil verwendete der Tisch sehr viel Zeit auf die Diskussion von Vorgehensweisen, Regeln und der Auslegung der Fragestellung. Das minderte die Effizienz und äußerte sich z. B. dadurch, dass diese Tischrunde länger arbeiten musste als die anderen Gruppen. Zudem waren die Tischergebnisse nicht so ausgefeilt, wie es sich die Teilnehmenden der Runde erhofft hatten.



Abb. 1 Screenshot der Arbeitsoberfläche des Online-Tisches. (Quelle: O. Scheel, ZIRIUS)

Online-Tisch-Diskussion war dadurch weniger immersiv, was in einer anderen Analysedimension der Studie eingehender betrachtet wurde. Analog zu Tischdiskussionen gab es im Online-Design zusätzliche Kanäle für den Austausch wie Emojis in einem Text-Chat. Eine ganze Anzahl von Moderationsinstrumenten unterstützte den Austausch am Online-Tisch, z. B. durch eine virtuelle Tischfläche, auf welcher die Teilnehmenden von ihnen beschriebene Klebezettel hängen konnten. Ein weiterer Kommunikationskanal des Online-Tisches war der Textchat, wo Teilnehmende parallel zum laufenden Austausch kurze Kommentare abgeben konnten. Online- wie Präsenz-Designs zeigten sich als flexibel, den laufenden Dialog an die Wünsche der Teilnehmenden anzupassen, z. B. in Bezug auf gemeinsame Regeln wie Entscheidungen getroffen werden sollen (Bütschi et al. 2004; Goldschmidt et al. 2012; Leach 2006; Stern et al. 1996).

5 Abschließende Reflexion von Ergebnissen und Vorgehen

Die inhaltsanalytische Betrachtung erbrachte folgende Erkenntnisse: Grundsätzlich ermöglichten alle untersuchten Designs den Austausch von Informationen, Argumenten und Perspektiven. Die Online-Foren unterschieden sich deutlich von allen Tischdesigns in Bezug auf die untersuchte Dimension der Intensität und Interaktivität. Die im einleitenden Teil konzeptuell abgeleiteten Kommunikationsmodi des deliberativen bzw. aggregierenden Austausches wurden in allen Designs festgestellt. Der deliberative Austausch ist dabei vergleichsweise intensiv. Dabei kann es durchaus deutliche Meinungs- und Perspektivunterschiede geben. Alle eingebrachten Inhalte und Argumente werden jedoch von Teilnehmenden gemeinsam reflektiert und im Regelfall in einem offenen Verständigungsprozess gemeinsam weiter ausgearbeitet bzw. in gemeinsam getragene Schlussfolgerungen überführt. Im aggregierenden Modus entwickeln die Teilnehmenden keinen intensiven Verständigungsprozess und damit auch keine gemeinsame Position. Bestimmte Meinungsaggregate bleiben bestehen. Diese konzeptuell und empirisch herausgestellten Kontraste sind wichtig. Besonders interessant ist jedoch die dritte Form des „additiven" Austausches, die sich bei den untersuchten Online-Foren auf Basis der inhaltsanalytischen Untersuchungen als typisch erwies. Wie die Auszählungen der Aktivitäten zeigten, äußerte sich der additive Modus durch niedrige Gesamtaktivitäten und durch eine betonte Lese- gegenüber der Schreibaktivität. Foren zeigten sich also weniger interaktiv bzw. die Rezeption von Informationen

dominierte. Intensive Reflexionen bzw. Erörterungen waren in den Foren eher selten. Auch der Charakter der Beiträge unterschied sich im Regelfall von den anderen Designs bzw. Austauschmodi. Die Forenbeiträge waren in der Regel kürzer als bspw. Beiträge in Tischdiskussionen. Im Extremfall verkürzte sich der Austausch bei Meinungsverschiedenheiten in den Foren auf das Teilen von Links zu Medienbeiträgen, worauf der Austausch abrupt endete. Der Charakter des additiven Stils in Foren mag oberflächlich betrachtet vor allem durch die technische Vermittlung geprägt erscheinen. Im Kontrast zu Foren bringen Teilnehmende in Tischdiskussionen seltener Beiträge z. B. in Form von Videos ein. Inputs dieser Art werden in Tischrunden in aller Regel von den entsprechenden Beitragenden erläutert bzw. dann in der Runde erörtert. An dieser Stelle kommen wir nun auf die Ausgangsfragen des Beitrags zurück.

Die Befunde zusammen genommen führen zu der Schlussfolgerung, dass die Qualität bzw. der Modus des Austauschs zuerst mit sozialen Faktoren in Zusammenhang steht. Technische Vermittlungsformen wie der Kontrast zwischen Online-Dialog oder Face-to-Face-Austausch stellen einen zwar wichtigen, aber nachgeordneten Faktor dar. Online-Foren verlieren mit den Befunden nicht an ihrem Wert für die Wissenschaftskommunikation. Im Gegenteil lässt sich ihr Wert nun differenzierter herausarbeiten. Sie unterstützen z. B. die Verbreitung von Informationen und ergänzten laufende Präsenz- bzw. Dialogformate wie im vorliegenden Fall zwei miteinander verbundene Bürger:innenkonferenzen. Es fanden sich Hinweise, dass Diskussionen bzw. Themen der Bürger:innenkonferenzen in Foren diskutiert wurden. Zudem erwiesen sich die Modi jedes Designs als nicht in Stein gemeißelt. Moderierende von Foren können mit gezielten Aktivitäten Threads bzw. Diskussionen in den Foren „beleben" und durchaus interaktiver machen, wie sich bei der „Idee des Monats" andeutete. Deliberative, von Angesicht zu Angesicht laufende Tisch-Diskussionen können wie Foren in andere Modi wechseln. Wenn Veranstaltungsdesigns auf einen deliberativen Austausch in einer Präsenzveranstaltung abzielen, eröffnet das viele Potenziale, aber ist noch kein Garant, dass der deliberative bzw. konstruktive Modus auch entwickelt und durchgehalten wird. Das deliberative Design ist also ebenfalls Anforderung und kein sicherer Hafen. Die explorative Untersuchung bestärkt zudem die grundlegende Schlussfolgerung aus der Hauptstudie, dass die Organisatoren von Dialog- und Beteiligungsverfahren die Wahl des Formats und nun auch der Designs, insbesondere hier Face-To-Face versus Online, unter Berücksichtigung der konkreten Zielsetzungen des geplanten Austauschprojektes bewusst abwägen sollten. Technik allein, also ohne ihre soziale Einbettung bedeutet also auch in Bezug auf die „Kanäle" in der Wissenschaftskommunikation weniger als man oberflächlich betrachtet annehmen könnte. Die Ergebnisse fordern geradezu weitere vertiefende Forschung zu den hier behandelten Fragestellungen.

Die inhaltsanalytische Betrachtung erbrachte folgende methodischen Erkenntnisse: Die vorliegende Studie stellt eine Exploration dar, welche Einbettung und Anwendung von inhaltanalytischen Betrachtungen im Multi-Method-Kontext des Formatvergleiches skizziert. Die Forschenden sehen sich bei diesen Multi-Method-Ansätzen in einem Spannungsverhältnis: Alles vergleichbar zu halten und gleichzeitig die Methoden passend zu den entsprechenden Kontexten einzusetzen. Vor allem in Bezug auf die Anpassung an die besonderen Erhebungskontexte bei Online-Designs erwiesen sich die inhaltsanalytischen Betrachtungen als zielführend. Dabei wurden die Gesamtstruktur der Argumente über die Argumentationsketten sowie strukturelle und inhaltliche Merkmale der Beiträge untersucht. Bei diesem Vorgehen zeichneten sich Analogien zur Methode der Beobachtung ab, die ebenfalls die Analyse der Häufigkeit von Redebeiträgen, aber auch von deren Struktur und Inhalt erlaubt. Beide Methoden treten so in Synergie, um Online- und Präsenzveranstaltungen vergleichend zu untersuchen. Argumentationsketten erwiesen sich als eine geeignete Analyseeinheit für die Format- und Design-übergreifende Beschreibung, Typisierung und Bewertung von Kommunikationsdesigns im Online- sowie Präsenzbereich.

Diese hier skizzierte integrative Vorgehensweise der vergleichenden Betrachtung von mehreren Dialogformaten bzw. -designs auf Basis mehrerer Methoden eröffnet Gestaltungsoptionen für die Begleitforschung in der Wissenschaftskommunikation. Ein großes Potenzial liegt in der *reflektierten* Annäherung an die Internetforschung. Zum Beispiel im Feld der Twitter-Forschung gibt es einen regelrechten Wettlauf um die Anzahl untersuchter Twitter-Nachrichten. Stichproben liegen hier häufig schon im Millionenbereich. Diese Forschung läuft hochgradig digitalisiert ab und nutzt sogenannte Dictionaries oder das Machine learning (Saif et al. 2016). Speziell die Sentimentanalyse zielt darauf ab, die mit Posts verbundenen subjektiven Wertungen zu untersuchen, was eine herausfordernde Aufgabe ist (Hussein 2016; Soleymani et al. 2017). Diese Analysen besitzen Stärken wie Schwächen. Eine differenzierte Studie, um Kommunikationen in Foren oder anderen Online-Designs strukturell und inhaltlich zu explorieren, zu beschreiben und zu typisieren, ist aufwendig. Sie verspricht jedoch hohes Potenzial, Forschungsmethoden weiter zu qualifizieren. Mit welcher Strategie eine Studie vorgeht, digitalisierend oder nicht bzw. ob mit kleinen oder riesigen Datenmengen, muss konzeptuell begründet und anhand des konkreten Forschungskontextes entschieden werden. Die Integration von inhaltsanalytischen Methoden in einen Multi-Methoden-Ansatz erwies sich als zielführend und verspricht Vorteile für aktuelle und zukünftige Forschungen im Feld der Wissenschaftskommunikation.

Literatur

Amichai-Hamburger Y, Gazit T, Bar-Ilan J, Perez O, Aharony N, Bronstein J, Dyne TS (2016) Psychological factors behind the lack of participation in online discussions. Comput Hum Behav 55:268–277

Bächtiger A, Dryzek JS, Mansbridge J, Warren ME (Hrsg) (2018) The Oxford Handbook of Deliberative Democracy. Oxford University Press

Baek YM, Wojcieszak M, Delli Carpini MX (2011) Online versus faceto-face deliberation: Who? Why? What? With what effects? New Media & Society 14:363–383. http://journals.sagepub.com/doi/pdf/https://doi.org/10.1177/1461444811413191. Zugriff am 13. Juli 2018

Barber B (1995) Participatory Democracy. In: Lipset SM (Hrsg) The encyclopedia of democracy. Routledge, London

Black LW (2009) Listening to the City: Difference, Identity, and Storytelling in Online Deliberative Groups. Journal of Public Deliberation 5(1):4. http://www.publicdeliberation.net/jpd/vol5/iss1/art4. Zugriff am 13. Juli 2018

Bobbio L (2010) Types of Deliberation. Journal of Public Deliberation 6(2):1. http://www.publicdeliberation.net/jpd/vol6/iss2/art1. Zugriff am 13. Juli 2018

Bütschi D, Carius R, Decker M, Gram S, Grunwald A, Machleidt P, Steyaert S, van Est R (2004) The practice of TA. Science interaction and communication. In: Decker M, Ladikas M (Hrsg) Bridges between science, society and policy. Technological assessment. Methods and impacts. Springer, Berlin, Heidelberg, New York, S 13–55

Carlitz R, Gunn R (2005) e-Rulemaking: a New Avenue for Public Engagement. Journal of Public Deliberation 1(1):7. http://www.publicdeliberation.net/jpd/vol1/iss1/art7. Zugriff 13. Juli 2018

Carpini MXD, Cook FL, Jacobs LR (2004) Public deliberation, discursive participation and citizen engagement: A review of the empirical literature. Annu Rev Polit Sci 7:315–344

Cohen J (1999a) Deliberation and democratic legitimacy. In: Bohman J, Rehg W (Hrsg) Deliberative democracy. Essays on reason and politics. Cambridge: MIT Press, S 67–92

Cohen J (1999b) Procedure and substance in deliberative democracy. In: Bohman J, Rehg W (Hrsg) Deliberative democracy. Essays on reason and politics. Cambridge: MIT Press, S 407–437

Denzin NK (2009) The research act. The theoretical introduction to sociological methods. Aldine Transaction, New Brunswick (US), London (UK)

Elster J (1999) The market and the forum: Three varieties of political theory. In: Bohman J, Rehg W (Hrsg) Deliberative democracy. Essays on reason and politics. Cambridge: MIT Press, S 3–33

Feindt P (2001) Regierung durch Diskussion? Peter Lang Verlag, Frankfurt, Diskurs- und Verhandlungsverfahren im Kontext von Demokratietheorie und Steuerungsdiskussion

Fuchs D (2007) Participatory, liberal and electronic democracy. In: Zittel T, Fuchs D (Hrsg) Participatory democracy and political participation. Can participatory engineering bring citizens back in? Routledge, Milton Park, S 29–53

Galston WA (1999) (How) does the internet effect the community? In: Kamarck EC, Nye JS Jr (Hrsg) Democracy.com. Governance in a networked world. Hollis Publishing, Hollis, NH, S 45–62

Gerhards J (1998) Öffentlichkeit. In: Jarren O, Sarcinelli U, Saxer U (Hrsg) Politische Kommunikation in der demokratischen Gesellschaft. Ein Handbuch mit Lexikonteil. Westdeutscher Verlag, Opladen, S 268–274

Gerhards J, Neidhardt F, Rucht D (1998) Zwischen Diskurs und Palaver. Westdeutscher Verlag, Opladen, Konturen der öffentlichen Meinungsbildung am Beispiel der deutschen Diskussion zur Abtreibung

Goldschmidt R (2014) Kriterien zur Evaluation von Dialog- und Beteiligungsverfahren. VS Verlag, Wiesbaden, Konzeptuelle Ausarbeitung eines integrativen Systems aus sechs Metakriterien

Goldschmidt R (2018) Selection of participatory formats as success factor for effective risk communication and decision-making processes. Conclusions from a systematic empirical format comparison. In: Journal of Risk Research 21(11):1331–1361. https://doi.org/10.1080/13669877.2017.1304975

Goldschmidt R, Scheel O, Renn O (2012) Zur Wirkung und Effektivität von Dialog- und Beteiligungsformaten. Universität Stuttgart (Stuttgarter Beiträge zur Risiko- und Nachhaltigkeitsforschung, Nr, Stuttgart, S 23

Goldschmidt R (2016) Wirkung von Beteiligungs- und Dialogverfahren: Der Vergleich von Formaten als Untersuchungsdesign. - Beispiel des Verbundprojektes „Wissenschaft debattieren!“ im Bereich der Wissenschaftskommunikation. In: Benighaus C, Wachinger G, Renn O (Hrsg) Bürgerbeteiligung. Konzepte und Lösungswege in der Praxis. Wolfgang Metzner Verlag, 309–328.

Habermas J (1998) Faktizität und Geltung. Suhrkamp Verlag, Frankfurt a. M, Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats

Hussein DMEM (2016) A survey on sentiment analysis challenges. Eng Sci 30:330–338

Jensen JL (2003) Public Spheres on the Internet: Anarchic or Government-Sponsored – A Comparison. Scand Polit Stud 26(4):349–374

Leach WD (2006) Public involvement in USDA Forest Service policymaking. A literature review. Journal of Forestry 104(1):43–49

Oliveira I, Tinoca L, Pereir A (2011) Online group work patterns: How to promote a successful collaboration. Comput Educ 57:1348–1357

Papadopoulos Y, Warin P (2007) Are innovative participatory and deliberative procedures in policy making democratic and effective? Eur J Polit Res 46(4):445–472

Quinlan S, Shephard M, Paterson L (2015) Online discussion and the 2014 Scottish independence referendum: Flaming keyboards or forums for deliberation? Elect Stud 38:192–205

Ravetz J (1998) The internet, virtual reality and real reality. In: Loader BD (Hrsg) Cyberspace Devide. Equality, agency and policy in the information age. London

Rosenfield I (1998) Internet - Die körperlose Psyche. In: Leggewie C, Maar C (Hrsg) Internet und Politik. Bollmann Verlag, Köln

Saif H, He Y, Fernandez M, Alani H (2016) Contextual semantics for sentiment analysis of Twitter. Inf Process Manage 52(1):5–19

Soleymani M, Garcia D, Jou B, Schuller B, Chang S, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14

Stern PC, Fineberg V (1996) Understanding risk: Informing decisions in a democratic society. National Academy Press (National Research Council, Committee on Risk Characterization), Washington, D. C

Stromer-Galley J, Bryant L, Bimber B (2015) Context and Medium Matter: Expressing Disagreements Online and Face-to-Face in Political Deliberations. Journal of Public Deliberation 11(1):1. http://www.publicdeliberation.net/jpd/vol11/iss1/art1. Zugriff 13 Juli 2018

Turkle S (1995) Live on the screen. Identity in the Age of the internet. Simon and Schuster, New York

Wilhelm AG (2000) Democracy in the digital age. Challenges to political life in cyberspace. Routledge, London

Wolfgang D, Jenkins J (2015) Diverse Discourse: Analyzing the Potential of Public Affairs Magazine Online Forums to Reflect Qualities of the Public Sphere. Journal of Public Deliberation 11(1):5. http://www.publicdeliberation.net/jpd/vol11/iss1/art5. Zugriff 13 Juli 2018

Wright S, Street J (2007) Democracy, deliberation and design: the case of online discussion forums. New Media Soc 9(5):849–869

Zittel T (2001) Elektronische Demokratie. Planskizze für die Demokratie des 21. Jahrhunderts. Neue Politische Literatur 46:433–470

Rüdiger Goldschmidt ist Soziologe und besitzt umfangreiche Expertise in quantitativer und qualitativer empirischer Sozialforschung sowie im Bereich der integrativen Analyseverfahren. Er verfügt über viel Erfahrung im Bereich Dialog und Beteiligung sowie Wissenschaftskommunikation und speziell mit Evaluationen von Initiativen bis hin zur weltweiten Ebene. Seine Tätigkeiten als Projektkoordinator und -leiter, Dozent und Moderator erfordern häufig interdisziplinäres und transdisziplinäres Denken und Vorgehen.

Oliver Scheel ist wissenschaftlicher Mitarbeiter und Projektleiter am Forschungsschwerpunkt ZIRIUS – Zentrum für interdisziplinäre Risiko- und Innovationsforschung der Universität Stuttgart sowie am HLRS – Höchstleistungsrechzentrum Stuttgart. Seine Forschungsschwerpunkte sind nachhaltige Techniknutzung, Energieversorgung und Energietechnik, diskursive sowie online-vermittelte Beteiligungserfahren sowie Wissenschaftskommunikation mit Fokus auf Sekundarschulen und Lehrkräfte.


cc
BY



Grundlagenbeitrag: Quantitative Testverfahren

Joachim Wirth und Jens Fleischer

Zusammenfassung

Quantitative Testverfahren kommen zum Einsatz, wenn Leistungen von Personen gemessen werden sollen. Im Rahmen der Evaluation von Wissenschaftskommunikation ist dies bspw. der Fall, wenn durch Wissenschaftskommunikation Wissen oder Fähigkeiten von Teilnehmenden verbessert werden sollen und Evaluation empirisch prüfen möchte, ob dieses Ziel erreicht wurde. Bei der Auswahl, Entwicklung und Bewertung von Testverfahren sind Gütekriterien einzuhalten, damit das empirische Ergebnis inhaltlich sinnvoll interpretierbare Aussagen zulässt. Die Bewertung der empirischen Ergebnisse erfolgt zudem vor dem Hintergrund einer Bezugsnorm, die bewusst gewählt werden muss. In dem Beitrag gehen wir auf die verschiedenen Testgütekriterien und Bezugsnormen ein und plädieren für interdisziplinäre Kooperationen bei der Auswahl und Entwicklung von quantitativen Testverfahren für die Evaluation von Wissenschaftskommunikation.

Eines der maßgeblichen Ziele von Wissenschaftskommunikation ist neben der Förderung von Interesse am wissenschaftlichen Thema, einer positiven Einstellung gegenüber dem Wert wissenschaftlicher Erkenntnisse oder auch des Vertrauens in die Wissenschaft – um nur einige Beispiele zu nennen – der Aufbau eines Wissenschaftsverständnisses (Bromme und Kienhues 2014).

J. Wirth (✉) · J. Fleischer
Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland
E-Mail: lehrlernforschung@rub.de

J. Fleischer
E-Mail: jens.fleischer@rub.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,*
https://doi.org/10.1007/978-3-658-39582-7_16

Expert:innen in einem bestimmten wissenschaftlichen Gebiet teilen ihr Wissen mit Personen, die in diesem Gebiet Lai:innen sind, sodass diese bspw. wissenschaftlich fundierte Entscheidungen in ihrem Alltag treffen können. Will man evaluieren, ob eine konkrete Maßnahme der Wissenschaftskommunikation bei den Adressat:innen tatsächlich zum gewünschten Wissenschaftsverständnis geführt hat, sollten Testverfahren im engeren Sinne (Rost 2004), kurz: Tests, zum Einsatz kommen. Tests in diesem engeren Sinne erfassen Leistungen. Sie grenzen sich damit von anderen Messverfahren, wie bspw. quantitativen Befragungen, ab, die keine Leistungen, sondern eher Selbsteinschätzungen verlangen. Wollte man bspw. evaluieren, ob die Besucher:innen eines Vortrags im Rahmen einer „Nacht der Wissenschaften" den Vortrag verstanden haben, könnte man am Ende des Vortrags die Besucher:innen bitten, einen Verständnistest zu bearbeiten, der eine Reihe von Frageitems mit vier Antwortoptionen enthält, wovon immer genau eine Antwortoption richtig ist. Die Summe der korrekt angekreuzten Antwortoptionen wäre dann ein quantitatives Maß zur Bestimmung des erworbenen Verständnisses. Ein solcher Test grenzte sich von einer Befragung ab, in der die Besucher:innen bspw. auf einer Rating-Skala von 1 (= „Ich habe nichts verstanden") bis 10 (= „Ich habe alles verstanden") ihr Verständnis selbst einschätzen sollen.

Im Folgenden beschäftigen wir uns mit Tests im Rahmen der Evaluation von Wissenschaftskommunikation. Einige der folgenden Ausführungen, bspw. zur Messung sogenannter latenter Merkmale oder zu den Testgütekriterien, sind auch für andere quantitative Messverfahren gültig. Wir werden sie jedoch immer mit Bezug auf Tests vorstellen.

1 Merkmale, Messen und Messverfahren

Tests haben wie alle quantitativen Messverfahren zum Ziel, die Ausprägung eines Personenmerkmals durch eine Zahl anzugeben. Dafür konstruieren sie eine Testskala, auf der diese Zahl abgelesen werden kann. Diese Testskala wird durch eine Reihe von Items gebildet. Personen bekommen für jedes Item eine vorab definierte Punktzahl gutgeschrieben in Abhängigkeit davon, wie erfolgreich sie das jeweilige Item bearbeitet haben. Besteht ein Wissenstest bspw. aus zehn Items, die korrekt oder nicht korrekt beantwortet werden können, und gibt jede korrekte Antwort genau einen Punkt, resultiert eine Skala von null bis zehn Punkten, wobei die Zahl Null die geringste und die Zahl Zehn die höchste messbare Ausprägung des Personenmerkmals Wissen anzeigt.

Skalen dieser Art können sich für die Messung unterschiedlichster Personenmerkmale eignen. Diese Personenmerkmale können dabei verschiedene Eigen-

schaften haben, die für ihre Messung methodisch relevant sind. Im Folgenden gehen wir auf drei dieser Eigenschaften etwas genauer ein.

1.1 Fähigkeiten

Personenmerkmale können dahingehend unterschieden werden, ob es sich bei ihnen um ein Leistungsmerkmal, also eine Fähigkeit, handelt oder nicht. Fähigkeiten können dabei verstanden werden als die sowohl kognitiven als auch motivationalen Voraussetzungen, die eine Person mitbringt, um in einem bestimmten Bereich Probleme lösen zu können (Weinert 2001). Für die Messung von Leistungsmerkmalen ergeben sich daraus u. a. zwei Konsequenzen. Erstens sind Leistungsmerkmale bereichsspezifisch. Expert:innen haben ihre Expertise in bestimmten Bereichen. Ein guter Schachspieler ist nicht automatisch gut in Monopoly, auch wenn es sich in beiden Fällen um Brettspiele handelt. Für die Messung von Leistungsmerkmalen bedeutet dies, dass die eingesetzten Messverfahren den betreffenden Bereich abdecken müssen und dieser dafür vorab genau definiert sein muss. Das mag trivial klingen, ist es aber meist nicht. Es erfordert eine systematische Analyse der Inhalte, die den Bereich definieren (z. B. Klauer 1987) sowie der Definition, auf welchem Niveau diese Inhalte verarbeitet bzw. die Leistungen gezeigt werden sollen (z. B. Anderson und Krathwohl 2001). So ist es bspw. eine andere (kognitive) Leistung, die Antwort auf eine Testfrage aus vier möglichen Antworten herauszusuchen und sie sozusagen „wiederzuerkennen“ als eine Antwort vollkommen frei selbst zu formulieren.

Zweitens sind Leistungsmerkmale nur Voraussetzungen. Das bedeutet, dass Personen nicht immer alle Probleme lösen, für deren Lösung sie die Voraussetzungen mitbringen. Manchmal realisieren sie in einer Situation gar nicht, dass sie über die Mittel zur Lösung des Problems verfügen, manchmal ist ihnen der Aufwand im Vergleich zum Nutzen zu hoch. Für die Messung von Leistungsmerkmalen bedeutet dies, dass die Messverfahren das entsprechende Leistungsmerkmal zunächst einmal aktivieren müssen, bspw. indem möglichst genau angegeben wird, welche Leistung bei der Bearbeitung der Items erwartet wird. Zum anderen muss bei der Durchführung des Messverfahrens gewährleistet sein, dass Personen bei der Bearbeitung der Items auch gewillt sind, Leistung zu zeigen. Das ist gerade bei der Evaluation von Maßnahmen der Wissenschaftskommunikation nicht immer einfach. Die Teilnahme an solchen Maßnahmen erfolgt freiwillig und meist in der Freizeit, sodass auch die Teilnahme an einer Leistungsmessung freiwillig ist, für die Freizeit investiert werden muss. Darüber hinaus ist die Messung von Leistungsmerkmalen für Personen immer mit dem

Risiko verbunden, eine geringe Leistung attestiert zu bekommen. Personen, die befürchten, geringe Leistungen zu zeigen, könnten in solchen Fällen zum self-handicapping (vgl. Schwinger und Stiensmeier-Pelster 2012) tendieren. Sie strengen sich nicht an und können dann eine geringe Leistung auf mangelnde Anstrengung zurückführen, ohne an der eigenen Fähigkeit zweifeln zu müssen. Für die Messung von Leistungsmerkmalen bedeutet dies, dass sie mit möglichst wenig zeitlichem oder sonstigen Aufwand verbunden sein und möglichst keine evaluative Funktion für die Personen haben sollte. Dafür sollte die Leistungsmessung vollständig anonymisiert durchgeführt werden, und es sollte keine Leistungsrückmeldung erfolgen.

1.2 Stabilität

Die meisten Leistungsmerkmale verändern sich über die Zeit hinweg, unterscheiden sich aber hinsichtlich der Dauer, die es benötigt, bis sie sich messbar verändert haben. So ist die Intelligenz einer Person bspw. ein vergleichsweise stabiles Personenmerkmal. Sicherlich entwickelt sich die Intelligenz über die Lebensspanne hinweg, messbare intraindividuelle Unterschiede im Abstand von wenigen Tagen, Wochen oder Monaten sind ab einem gewissen Lebensalter jedoch nicht mehr zu erwarten. Anders verhält es sich bspw. mit Wissensständen. Allein die Partizipation an einem Citizen-Science-Projekt kann innerhalb weniger Stunden zu einer messbaren Veränderung von Wissen im entsprechenden Bereich führen.

Im Rahmen der Evaluation von Wissenschaftskommunikation sind veränderbare Leistungsmerkmale selbstredend die interessanteren. Wenn ein Ziel von Wissenschaftskommunikation der Aufbau von Wissenschaftsverständnis ist, dann ergibt dieses Ziel nur unter der Voraussetzung Sinn, dass das Wissenschaftsverständnis innerhalb der Zeitspanne, die die Maßnahme andauert, veränderbar ist. Stabile Personenmerkmale können jedoch im Rahmen einer Evaluation durchaus auch interessant sein, bspw. um Randbedingungen zu erfassen, die gegeben sein müssen, damit eine Maßnahme der Wissenschaftskommunikation das Wissenschaftsverständnis beeinflussen kann.

Die Stabilität eines Personenmerkmals ist jedoch nicht nur inhaltlich relevant, sondern hat auch bestimmte methodische Konsequenzen. So ist der Zeitpunkt, zu dem ein Personenmerkmal erfasst wird, bei veränderbaren Merkmalen natürlich relevant, während stabile Merkmale nahezu jederzeit gemessen werden können. Abgesehen vom Zeitpunkt ist auch die Häufigkeit der Messungen zwischen

stabilen und veränderbaren Personenmerkmalen unterschiedlich. Kommt man bei stabilen Merkmalen mit einer Messung aus, lässt sich die (oftmals gewünschte) Veränderung eines Personenmerkmals nur durch wiederholte Messung, bspw. in einem Prä-Post-Testdesign, bei dem das interessierende Merkmal vor und nach einer Maßnahme erfasst wird, messen. Dies mag trivial klingen, wird bei der Evaluation von Wissenschaftskommunikation jedoch erstaunlicherweise häufig versäumt (z. B. Masters et al. 2016).

1.3 Beobachtbarkeit

Als letzte Eigenschaft zur Kategorisierung von Personenmerkmalen gehen wir auf die Beobachtbarkeit eines Merkmals ein. Personenmerkmale können direkt beobachtbar sein oder nicht. Im ersten Fall spricht man von einem manifesten, im letzten von einem latenten Merkmal. Merkmale wie das Alter oder das Monatsgehalt sind direkt beobachtbar und damit manifest. Sie lassen sich (zumindest theoretisch) mit einem Blick auf den Personalausweis oder auf einen Kontoauszug in ihrer Ausprägung messen. Andere Merkmale wie das Wissen über ein Thema lassen sich nicht direkt beobachten und sind damit latent. Wie viel eine Person über ein Thema weiß, kann man ihr nicht direkt ansehen. Da es sich bei Messverfahren jedoch um empirische, sprich auf Wahrnehmung basierende Verfahren handelt, ist die Messung nicht direkt beobachtbarer Merkmale natürlich eine Herausforderung. Um dieser zu begegnen, bedienen sich Messverfahren einer Hilfskonstruktion. Sie definieren ein Verhalten, das zwei Eigenschaften hat. Zum einen muss das definierte Verhalten direkt beobachtbar sein. Zum anderen muss es vom eigentlich zu messenden latenten Personenmerkmal maßgeblich beeinflusst werden.

Leistungsmerkmale sind latente Personenmerkmale. Ihre Ausprägung wird mithilfe von Tests erfasst, bei denen Personen sich im Rahmen von Testitems verhalten müssen. Dieses Verhalten wird beobachtet und in Abhängigkeit der Ausprägung des beobachteten Verhaltens erhält die Person eine Zahl auf der Testskala. Damit das alles funktioniert, müssen einige Kriterien erfüllt sein, die gewährleisten sollen, dass das beobachtete Verhalten und seine Übersetzung in eine Zahl möglichst ausschließlich von dem interessierenden latenten Leistungsmerkmal bestimmt sind. Nur in diesem Fall kann von einer hinreichenden Testgüte ausgegangen werden. Die entsprechenden Kriterien werden daher auch Testgütekriterien genannt.

2 Testgütekriterien

Alle guten Tests erfüllen mindestens drei Hauptgütekriterien, zu denen die Reliabilität, die Objektivität und die Validität gezählt wird (siehe auch Böhmer und Abacioglu in diesem Band). Diese drei Gütekriterien haben gemeinsam, dass sie alle den Schutz des zu beobachtenden Verhaltens und seiner Übersetzung in eine Zahl vor ungewollten Einflüssen beschreiben. Ungewollte Einflüsse sind dabei solche, die nicht von dem zu messenden latenten Leistungsmerkmal ausgehen. Sie können systematisch oder unsystematisch sein.

2.1 Reliabilität

Unsystematische Einflüsse sind Gegenstand der sogenannten „Klassischen Testtheorie" (Gulliksen 1950; siehe auch Döring und Bortz 2016) und das Gütekriterium, das sich mit dem Schutz vor solchen unsystematischen Einflüssen beschäftigt, wird Reliabilität (= Zuverlässigkeit oder Fehlerfreiheit) genannt. Die klassische Testtheorie beschäftigt sich mit ungewollten Einflüssen auf das beobachtete Verhalten oder seiner Übersetzung in eine Zahl, die zufällig (und damit unsystematisch) in zweierlei Hinsicht sind. Erstens ist das Ausmaß des Einflusses rein zufällig. Zweitens ist es völlig zufällig, ob durch den jeweiligen Einfluss das Ausmaß des latenten Leistungsmerkmals über- oder unterschätzt wird. Ein Beispiel für solche zufälligen Einflüsse wäre das Ankreuzen der richtigen Antwort in einem Wissenstest, obwohl das entsprechende Wissen gar nicht verfügbar ist, wenn einer Person also durch bloßes Raten ein „Glückstreffer" gelingt. Die Summe aller zufälligen Einflüsse wird in der klassischen Testtheorie unter dem Begriff des (Mess-)Fehlers zusammengefasst, weshalb sie auch als Messfehlertheorie bekannt ist. Grundlage dieser Theorie ist die Annahme, dass der Messwert, also die Zahl, durch die das Ausmaß des latenten Personenmerkmals ausgedrückt werden soll, sich zusammensetzt aus dem „wahren" Wert und dem Messfehler. Der wahre Wert ist dabei die Zahl, die zustande käme, wenn es keinerlei zufällige Einflüsse gäbe. Der Messfehler ist der (positive oder negative) Betrag, der auf die Summe aller (unbekannten) zufälligen Einflüsse zurückzuführen ist. Die Frage der Reliabilität ist nun, wie hoch der Anteil des wahren Werts sowie der des Messfehlers am Messwert ist. Ist der Anteil des wahren Werts sehr hoch und damit der Anteil des Messfehlers sehr niedrig, ist der Messwert sehr reliabel und zuverlässig.

Sowohl für die Gewährleistung als auch für die Überprüfung der Reliabilität eines Tests ist die Wiederholung von Messungen notwendig. Das ist der Grund, wieso Tests nicht aus einem, sondern aus mehreren Items bestehen. Jedes Item stellt dabei eine Messung dar. Unter der Voraussetzung, dass das Verhalten, das bei der Bearbeitung aller Items eines Tests beobachtet wird, von demselben latenten Personenmerkmal beeinflusst wird, wird durch die Bearbeitung mehrerer Items die Messung des latenten Personenmerkmals mehrfach wiederholt. Werden diese wiederholten Messungen dann zusammengefasst, indem man bspw. die Punkte, die bei den Items jeweils erreicht wurden, aufsummiert, resultiert ein reliables Testergebnis. Der Grundgedanke hierbei ist der folgende: Der Messwert eines jeden Items setzt sich zusammen aus „wahrem" Wert und Messfehler. Die wahren Werte weisen alle mehr oder weniger ausgeprägte, aber immer positive Beträge auf. Die Messfehler haben dagegen jeweils zufälligerweise einen mehr oder weniger ausgeprägten positiven oder negativen Betrag. Summiert man die Messwerte der Items, summiert man sowohl die wahren Werte als auch die Messfehler. Da alle wahren Werte positiv sind, ist auch ihre Summe positiv. Da im Gegensatz dazu die Messfehler sowohl positive als auch negative Beträge aufweisen, strebt ihre Summe gegen Null. Die Summe der Messwerte, also das Testergebnis, ist dadurch maßgeblich von der Summe der wahren Werte bestimmt, während die Summe der Messfehler einen gegen Null strebenden Anteil daran hat. Aus dieser Überlegung heraus folgt, dass ein Test umso reliabler ist, aus je mehr Items er besteht, deren Bearbeitung von demselben latenten Personenmerkmal maßgeblich beeinflusst wird. Durch die Wiederholung der Messung und dem Zusammenfassen der Messergebnisse wird die Reliabilität des Testergebnisses gewährleistet.

So zumindest die (klassische Test-)Theorie. Ob ein Testergebnis aber nicht nur theoretisch, sondern auch tatsächlich reliabel ist, lässt sich ebenfalls durch Wiederholung überprüfen. Dabei werden verschiedene Arten der Wiederholung unterschieden (Döring und Bortz 2016), bspw. in Abhängigkeit von der Stabilität des zu messenden Leistungsmerkmals. Handelt es sich um ein stabiles Leistungsmerkmal, lässt sich die Reliabilität eines Tests als Test-Retest-Reliabilität bestimmen. Der Gedanke dabei ist, dass man dieselben Personen denselben Test nicht nur einmal, sondern mit einem gewissen zeitlichen Abstand mindestens zweimal bearbeiten lässt. Ist das latente Personenmerkmal stabil, sollten diejenigen, die beim ersten Bearbeiten ein gutes Testergebnis erzielt haben, beim zweiten Bearbeiten ebenfalls ein gutes Testergebnis erreichen. Ist ein Test reliabel und ist das getestete Personenmerkmal stabil, dann stehen die Testergebnisse bei einer wiederholten Testung in einem systematischen Zusammenhang. Ein solcher systematischer Zusammenhang lässt sich als Korrelation berechnen. Ein Kor-

relationskoeffizient mit einem hohen positiven Betrag ($r > 0{,}8$) drückt dann eine hohe Test-Retest-Reliabilität aus.

Ist eine wiederholte Testdurchführung nicht sinnvoll, bspw. weil das zu erfassende Personenmerkmal veränderbar ist, besteht für die Überprüfung der Reliabilität die Möglichkeit, die Wiederholung nicht auf der Ebene des gesamten Tests bzw. der Testskala anzusiedeln, sondern auf der Ebene der einzelnen Items eines Tests, welche die Skala letztlich bilden. Der Gedanke ist, dass eine Person, die ein Testitem gut oder korrekt bearbeitet hat, mit hoher Wahrscheinlichkeit auch die weiteren Testitems gut oder korrekt bearbeiten wird. Ist ein Test reliabel, dann stehen die Messergebnisse der einzelnen Testitems in einem systematischen Zusammenhang. Die Testitems bilden in diesem Fall eine intern konsistente Testskala. Das Ausmaß dieser internen Konsistenz gilt als Schätzer der Reliabilität eines Tests und lässt sich über Koeffizienten wie bspw. Cronbachs Alpha (Cronbach 1951) ausdrücken. Auch hier gilt, dass ein hoher positiver Betrag (Cronbachs Alpha > 0,8) eine hohe interne Konsistenz und damit eine hohe Testreliabilität ausdrückt.

2.2 Schwierigkeit

Findet man starke, systematische Zusammenhänge zwischen den Messwerten der Items einer Testskala oder zwischen den Ergebnissen eines wiederholt durchgeführten Tests, dann ist das ein Indikator für die hohe Reliabilität des Tests. Sind diese Zusammenhänge dagegen niedrig ausgeprägt, kann eine geringe Reliabilität die Ursache sein, muss aber nicht. Es kann auch sein, dass der Test für die untersuchten Personen viel zu leicht oder viel zu schwierig ist. In dem Fall fallen die berechneten Koeffizienten niedrig aus und unterschätzen die Reliabilität des Tests. Will man bspw. evaluieren, ob eine Maßnahme der Wissenschaftskommunikation zu einem umfangreicheren Wissen führt, ist man gut beraten, bei den betreffenden Personen das Wissen sowohl vor als auch nach der Maßnahme zu testen. Wenn die Personen jedoch mit wenig Vorwissen in die Maßnahme starten, können sie beim Vorwissenstest bei den meisten Items wahrscheinlich nur raten. Dieses Raten führt zu unsystematischen, zufälligen Antworten. Die Messergebnisse der einzelnen Testitems werden stark vom Zufall geprägt sein und die interne Konsistenz der Testskala entsprechend niedrig ausfallen. Das muss jedoch nicht bedeuten, dass die Skala bzw. der Test an sich wenig reliabel ist. Es ist sehr gut möglich, dass derselbe Wissenstest, nach der Maßnahme der Wissenschaftskommunikation eingesetzt, zu intern konsistenten Ergebnissen führt, da die Personen jetzt über ein mehr oder weniger ausgeprägtes Wissen verfügen und

entsprechend bei mehr oder weniger vielen Items die korrekte Antwort kennen (und nicht raten). In dem Fall ist die Schwierigkeit des Tests für die Gruppe der untersuchten Personen angemessen, und die berechneten Koeffizienten sind gute Schätzer der Reliabilität.

Während sich die klassische Testtheorie maßgeblich mit der Frage des Messfehlers beschäftigt, ist das Zusammenspiel von Ausprägung eines latenten Personenmerkmals und Itemschwierigkeit zentraler Gegenstand einer anderen Testtheorie, der sogenannten Item-Response-Theory (kurz IRT; z. B. van der Linden 2016). Die IRT beschäftigt sich maßgeblich mit der gleichzeitigen Schätzung sogenannter Personenparameter (= Ausprägung des latenten Personenmerkmals) und sogenannter Itemparameter (= Schwierigkeit eines Items). Tests, die gemäß der IRT konstruiert wurden, erlauben die Schätzung des Personenmerkmals unabhängig von der Itemschwierigkeit sowie die Schätzung der Itemschwierigkeit unabhängig von der Ausprägung des latenten Personenmerkmals bei den getesteten Personen. Dies hat natürlich Vorteile für Testsituationen wie im oben beschrieben Fall der Testung von Vorwissen. Die IRT wird aber auch herangezogen, wenn bspw. ein Test mehr Items enthält als eine Person zu einem Zeitpunkt bearbeiten könnte. IRT-skalierte Tests erlauben es in diesem Fall, Personen nur eine Auswahl der Testitems bearbeiten zu lassen und trotzdem zu vergleichbaren Messungen des latenten Personenmerkmals zwischen verschiedenen Personen zu kommen. Die IRT kommt zudem zum Einsatz, wenn adaptive Tests konstruiert werden sollen, bei denen Personen nur eine Itemauswahl präsentiert bekommen, die individuell an die jeweilige Ausprägung ihres Personenmerkmals angepasst wird. Auch dann kommen solche Tests zu vergleichbaren Ergebnissen zwischen Personen, obwohl die Personen unterschiedliche Items bearbeitet haben. Derartige Möglichkeiten bieten Tests, die auf Basis der klassischen Testtheorie konstruiert wurden, nicht.

Im Rahmen der IRT werden Personen- und Itemparameter auf derselben Skala abgetragen. Dadurch können Personen- und Itemparameter direkt miteinander verglichen werden, was eine kriteriumsorientierte (inhaltliche) Interpretation der Testergebnisse ermöglicht (Embretson und Reise 2000). Anstatt das Testergebnis ausschließlich in Form einer Zahl zu erhalten, können so Wertebereiche auf der Testskala definiert werden. Und diese Bereiche können inhaltlich durch die (Teil-) Leistungen beschrieben werden, die eine Person (mindestens) zeigen kann, wenn sie Items aus diesem Wertebereich mit hinreichender Wahrscheinlichkeit erfolgreich bearbeiten kann. Dadurch wird das Testergebnis inhaltlich interpretierbar. Bekannt geworden sind solche IRT-basierten Interpretationen bspw. im Kontext der Schulleistungsvergleichsstudien wie PISA, in denen Testergebnisse in Form von Kompetenzniveaus rückgemeldet werden. Diese stellen eine qualitative,

kriteriumsorientierte Beschreibung der Anforderungen dar, die Personen mit entsprechend ausgeprägten Kompetenzen bewältigen können (Hartig 2007; Reiss et al. 2016).

Die Entwicklung geeigneter Testverfahren und die Auswertung der durch sie gewonnenen Daten auf Basis von IRT-Modellen stellt mitunter höhere Anforderungen an die methodische Expertise von Wissenschaftler:innen und ist insgesamt auch mit einem höheren Aufwand verbunden als dies bei der klassischen Testtheorie der Fall ist. Es stehen inzwischen jedoch zahlreiche anwender:innenfreundliche Softwarepakete zur Verfügung, die einen breiten Einsatz von IRT-Modellen ermöglichen (Kelava und Moosbrugger 2020). Eine maßgebliche Einschränkung für den Einsatz von IRT-Modellen gegenüber der klassischen Testtheorie stellen allerdings die recht hohen Anforderungen an die benötigten Personenstichproben dar, die mit zunehmender Komplexität der Modelle steigen (de Ayala 2009).

2.3 Objektivität

Bislang haben wir uns mit unsystematischen Einflüssen auf das in einem Test gezeigte Verhalten und seiner Übersetzung in eine Zahl beschäftigt. In den folgenden zwei Unterkapiteln widmen wir uns systematischen Einflüssen, die aber gleichermaßen ungewollt das Verhalten beeinflussen. Diese Einflüsse können ihren Ursprung entweder innerhalb der testenden Person oder innerhalb der getesteten Person haben.

Wenden wir uns zunächst dem ersten Fall zu. Der Schutz vor ungewollten, systematischen Einflüssen, die von der testenden Person ausgehen, betrifft das Gütekriterium der Objektivität. Die Objektivität eines Tests ist dann gefährdet, wenn die testdurchführende Person Einfluss auf das Verhalten der getesteten Person oder auf dessen Bewertung nimmt. Dies muss gar nicht absichtlich geschehen. Häufige Ursache für unzureichender Testobjektivität sind mangelhafte Anweisungen für die Testdurchführung. Wenn bspw. keine Angabe dazu besteht, wie lange Personen Zeit haben, einen Test zu bearbeiten, müssen Testdurchführende selbst einschätzen, wann die Testbearbeitung beendet werden soll, was zu ungewollten Unterschieden in der Testdurchführung führt. Eine andere übliche Quelle für mangelnde Objektivität sind unzureichende Bewertungsschlüssel im Falle von Testitems mit einem offenen Antwortformat. Wenn bspw. ein Frageitem durch einen selbstständig zu formulierenden Text beantwortet werden soll, muss

im Nachhinein der Antworttext in Bezug auf die darin ausgedrückte Leistung interpretiert werden. Wenn für diese Interpretation keine sehr klaren Regeln und Kriterien vorab definiert wurden, wird derselbe Antworttext bei verschiedenen Interpretierenden zu unterschiedlichen Leistungseinschätzungen, sprich Messergebnissen führen. Vor diesem Hintergrund sind Items mit einem geschlossenen Antwortformat, also Items, die durch das Ankreuzen einer Option oder das Ausfüllen durch genau ein Wort oder genau eine Zahl zu beantworten sind, natürlich zu bevorzugen, da bei ihnen kein Interpretationsspielraum und damit ein Höchstmaß an Objektivität gegeben ist.

Um eine hinreichende Objektivität zu gewährleisten, haben gute Testverfahren entsprechende Testmanuale, in denen sehr genau beschrieben ist, 1) wie ein Test durchzuführen ist, 2) wie das beobachtete Verhalten auszuwerten ist (also welchem Verhalten welche Zahl zuzuordnen ist) und 3) wie genau das resultierende Testergebnis zu interpretieren ist (bspw. ob ein Testergebnis ein eher durchschnittliches oder ein über- oder unterdurchschnittliches Ergebnis darstellt). Der Sinn dieser Manuale ist, den Interpretationsspielraum der Testdurchführenden, häufig Rater (engl. to rate = bewerten, beurteilen) genannt, in allen Phasen der Testdurchführung, -auswertung und -interpretation möglichst eng zu gestalten. Ob dies hinreichend gelungen ist, kann dadurch überprüft werden, dass nicht nur ein Rater einen Test bei derselben Gruppe von Personen durchführt, sondern mindestens zwei. Das Prinzip ist dasselbe wie im Falle der Überprüfung der Reliabilität. Wieder wird die Messung wiederholt, nur dieses Mal nicht zu mehreren Testzeitpunkten oder durch mehrere Items, sondern durch mehrere Rater. Und wie bei der Reliabilität wird auch zur Überprüfung der Objektivität die Stärke eines Zusammenhangs geprüft. Ist ein Testverfahren objektiv, sollten Personen, die von einem:r Rater:in eine hohe Leistung attestiert bekommen, auch von einem:r anderen Rater:in eine hohe Leistung zugesprochen bekommen. Auch ein solcher systematischer Zusammenhang zwischen den Ergebnissen zweier oder mehrerer Rater:innen lässt sich als Korrelation berechnen, die ein sogenanntes Interrater-Agreement ausdrückt. Ist das Interrater-Agreement hoch, ist das Kriterium der Objektivität mit hoher Wahrscheinlichkeit erfüllt. In Abhängigkeit vom Skalenniveau der Testdaten stehen zur Berechnung des Interrater-Agreements verschiedene Koeffizienten zur Verfügung. Der wohl bekannteste, aber auch recht strenge Koeffizient ist Cohens Kappa. Bei Werten von Cohens Kappa > 0,8 geht man von einem hohen Interrater-Agreement und damit von einer hinreichenden Objektivität aus.

2.4 Validität

Mit der Validität ist ein Testgütekriterium angesprochen, bei dem systematische und ungewollte Einflüsse ihren Ursprung innerhalb der getesteten Person haben. Solche Einflüsse kommen von Personenmerkmalen, die ein anderes Merkmal als das eigentlich interessierende latente Personenmerkmal sind. Wenn bspw. ein Wissenstest nach einer Wissenschaftskommunikationsmaßnahme eingesetzt wird, dessen Items lange Sätze mit komplexer Satzstruktur und eine Fülle an Fremdwörtern beinhalten, dann besteht die Gefahr, dass mit diesem Test weniger das Verständnis des kommunizierten Wissens als vielmehr Lese- und Sprachfähigkeiten gemessen werden. Ein weiteres, leider häufig vorkommendes Beispiel, wurde oben bereits beschrieben. Es betrifft den Einsatz von Befragungen mit Rating-Skalen zur Erfassung von Wissen oder Verständnis (z. B. Land-Zandstra et al. 2016). Auch hier muss man sich im Klaren sein, dass nicht das Wissen oder das Verständnis selbst damit erfasst werden, sondern die selbsteingeschätzte Bewertung derselben. Diese Bewertung kann in hohem Maße beeinflusst sein durch das entsprechende Wissen oder Verständnis. Sie kann aber auch stark von der Fähigkeit, sich selbst einzuschätzen oder dem Drang, sozial erwünscht zu antworten, abhängen.

Sowohl für die Gewährleistung der Validität bei der Konstruktion oder der Auswahl von Tests, als auch für die Überprüfung ihrer Validität ist ein fundiertes theoretisches Wissen über das zu messende latente Personenmerkmal unabdingbar. Nur wenn das Personenmerkmal genau theoretisch beschrieben und definiert ist, kann die Validität eines Tests eingeschätzt werden. Dies erfordert entsprechende, meist psychologische oder fachdidaktische Expertise in dem Bereich, dem das latente Personenmerkmal zuzuordnen ist. Sollte man selbst eher Lai:in in diesem Gebiet sein, ist man gut beraten, eine:n entsprechende:n Expert:in um Hilfe zu bitten und sich nicht auf das eigene womöglich lai:innenhafte Alltagsverständnis zu verlassen.

Um zu überprüfen, ob ein eingesetzter Test zu validen Ergebnissen führt, prüft man, wie bereits bei der Reliabilität und der Objektivität, in der Regel Zusammenhänge; dieses Mal jedoch nicht zwischen Messzeitpunkten, Items oder Ratern, sondern zwischen verschiedenen Personenmerkmalen (Cronbach und Meehl 1955). Auch dafür sind fundierte Kenntnisse des aktuellen Forschungsstandes in Bezug auf das interessierende latente Personenmerkmal notwendig. Auf deren Basis wird entschieden, mit welchen Personenmerkmalen das zu testende latente Personenmerkmal in einem engen Zusammenhang steht („konvergente Validität") und mit welchen Personenmerkmalen kein Zusammenhang besteht („diskriminante Validität"). Diese auf der theoretischen Ebene

postulierten (Nicht-)Zusammenhänge werden dann mithilfe des zu validierenden Tests empirisch überprüft. Entsprechen die empirisch ermittelten Korrelationen den theoretischen, auf dem aktuellen Stand der Forschung abgeleiteten Zusammenhängen, sprechen diese Zusammenhänge für die Validität des Tests. Dabei gelten Korrelationen von $r > 0{,}6$ im Falle konvergenter Zusammenhänge als hoher Validitätskoeffizient (Weise 1975), im Falle diskriminanter Zusammenhänge sollte r sich nicht bedeutsam von Null unterscheiden.

3 Bezugsnormen

Das Ergebnis eines Tests ist eine Zahl. Diese Zahl ist zunächst einmal bedeutungslos. Was bedeutet es bspw. wenn jemand in einem Wissenstest 7 von 10 Punkten erreicht hat? Verfügt die Person dann über viel oder wenig Wissen? Auf Grundlage der bloßen Zahl lässt sich diese Frage nicht beantworten, es sei denn, es handelt sich um einen IRT-basierten Test, für den Testwertebereiche inhaltlich beschrieben wurden. In den meisten Fällen haben wir es aber mit klassisch konstruierten Tests zu tun, und deren Ergebnisse lassen sich nur mithilfe von Vergleichsmaßstäben bewerten.

Diese Vergleichsmaßstäbe werden in der Literatur als Bezugsnormen bezeichnet. Unterschieden werden dabei die soziale, die individuelle sowie die kriteriale Bezugsnorm (Heckhausen 1974). Wendet man eine soziale Bezugsnorm an, vergleicht man ein Testergebnis mit den Testergebnissen anderer Personen. Über diesen Vergleich lässt sich einschätzen, ob ein Testergebnis besser, schlechter oder ähnlich wie die Testergebnisse anderer ist. Die soziale Bezugsnorm kommt im Rahmen der Evaluation von Wissenschaftskommunikation bspw. zum Einsatz, wenn in einer Evaluationsstudie ein experimentelles Design mit einer Interventions- und einer Kontrollgruppe realisiert wurde. Angenommen man wollte evaluieren, ob die Teilnahme an einem Schülerlaborprojekt aus der Chemie bei den Schüler:innen zu Kenntnissen naturwissenschaftlicher Arbeitsweisen führt, könnte man die Schüler:innen dieser Interventionsgruppe am Ende des Schülerlaborprojekts einen Test zu naturwissenschaftlichen Arbeitsweisen (Klos et al. 2008) bearbeiten lassen. Denselben Test würde man aber auch Schüler:innen geben, die „nur“ den herkömmlichen Chemieunterricht besuchten, und damit die Kontrollgruppe bildeten. Auch wenn dieses ein immer noch recht schwaches Untersuchungsdesign wäre, könnte man die durchschnittliche Testleistung der Interventionsgruppe mit der entsprechenden Testleistung der Kontrollgruppe vergleichen. Das Schülerlaborprojekt würde dann als erfolgreich

bewertet, wenn die durchschnittliche Testleistung in der Interventionsgruppe höher ausfiele als die der Kontrollgruppe.

Bei der individuellen Bezugsnorm werden die Testleistungen einer Person mit den Leistungen derselben Person im selben Test zu einem früheren Zeitpunkt verglichen. Die individuelle Bezugsnorm kommt also zum Tragen, wenn Veränderungsmessungen durchgeführt werden. Ob eine Testleistung als mehr oder weniger gut bewertet wird, hängt in diesem Fall davon ab, ob eine Testleistung besser oder schlechter als frühere Testleistungen derselben Person ist. Die individuelle Bezugsnorm kommt im Rahmen der Evaluation von Wissenschaftskommunikation bspw. zum Einsatz, wenn in einer Evaluationsstudie ein Prä-Post-Testdesign realisiert wurde. Wollte man bspw. evaluieren, ob die Teilnahme an einem Schülerlaborprojekt aus der Chemie bei den Schüler:innen zu Kenntnissen naturwissenschaftlicher Arbeitsweisen führt, könnte man die Schüler:innen nicht nur nach der Teilnahme am Schülerlaborprojekt, sondern zusätzlich auch davor einen Test zu naturwissenschaftlichen Arbeitsweisen (Klos et al. 2008) bearbeiten lassen. Das Schülerlaborprojekt würde dann als erfolgreich bewertet, wenn die Testleistungen nach der Teilnahme bedeutsam höher ausfielen als vor der Teilnahme, obwohl auch bei diesem Design angemerkt werden muss, dass es sich um ein schwaches Untersuchungsdesign handelt. Ideal wäre ein Untersuchungsdesign, das ein experimentelles Design und ein Prä-Post-Testdesign kombiniert, wenn also die Tests bei sowohl einer Interventionsgruppe, als auch einer Kontrollgruppe sowohl vor einer Intervention als auch nach einer Intervention eingesetzt würden. In dem Fall wäre eine Kombination der sozialen und der individuellen Bezugsnorm möglich und man würde das Ausmaß der Veränderungen in den beide Gruppen miteinander vergleichen. Ein Schülerlaborprojekt würde in dem Fall dann als erfolgreich bewertet, wenn die teilnehmenden Schüler:innen mehr hinzulernten als die Schüler:innen, die „nur" herkömmlichen Chemieunterricht genossen haben.

Das Heranziehen der individuellen Bezugsnorm ist nur unter bestimmten Bedingungen sinnvoll. Zum einen darf das zu messende Personenmerkmal kein stabiles sein. Zum anderen muss ausgeschlossen werden, dass sich Personen bei der wiederholten Testbearbeitung an vorherige Bearbeitungen so erinnern, dass sie daraus Vorteile bei der erneuten Bearbeitung haben. Hier ist bspw. die Zeitspanne zwischen den Testungen von entscheidender Bedeutung.

Die kriteriale Bezugsnorm bewertet ein Testergebnis unabhängig von Testleistungen anderer Personen oder anderer Testzeitpunkte. Für die kriteriale Bezugsnorm werden vorab Testwerte definiert, die erreicht werden müssen, damit eine Testleistung als mehr oder weniger gut bewertet wird. Im Falle von IRT-basierten Tests sind das die angesprochenen inhaltlich beschreibbaren Werte-

bereiche. Bei klassisch konstruierten Tests sind das einzelne Werte auf der Testskala, die mindestens erreicht werden müssen, damit ein Testergebnis als gut bewertet wird. Angenommen man wollte evaluieren, ob die Teilnahme an einem Schülerlaborprojekt aus der Chemie bei den Schüler:innen zu hinreichenden Kenntnissen naturwissenschaftlicher Arbeitsweisen führt, dann könnte man die Schüler:innen nach der Teilnahme am Schülerlaborprojekt einen Test zu naturwissenschaftlichen Arbeitsweisen (Klos et al. 2008) bearbeiten lassen. Zudem würde man vorab definieren, dass eine hinreichende Testleistung bei mindestens 70 % korrekt beantworteter Items liegt. Das Schülerlaborprojekt würde dann als erfolgreich bewertet, wenn die durchschnittliche Testleistung nach der Teilnahme bei 70 % oder höher läge.

4 Fazit

Tests kommen im Rahmen der Evaluation von Maßnahmen der Wissenschaftskommunikation immer dann zum Einsatz, wenn Personenmerkmale erfasst werden sollen, die Voraussetzungen für Leistungen sind. Damit die Ausprägung dieser meist nicht direkt beobachtbaren Personenmerkmale durch einen Testwert gut eingeschätzt werden kann, muss der Test Gütekriterien erfüllen, wozu insbesondere die Reliabilität, die Objektivität und die Validität zählen. Das Erfüllen dieser Kriterien gewährleistet, dass das im Test beobachtete Antwortverhalten maßgeblich vom interessierenden Personenmerkmal und nicht von anderen, teilweise unbekannten Faktoren beeinflusst wird. Die Entwicklung solcher Tests sowie die empirische Überprüfung der Gütekriterien ist mit großem Aufwand verbunden, in die inhaltliche und methodische Expertise, Zeit und nicht zuletzt Geld investiert werden muss. Vor diesem Hintergrund ist man gut beraten, auf bereits entwickelte und empirisch bewährte Tests zurückzugreifen.

Doch auch die Suche und Auswahl solcher etablierter Testverfahren ist nicht ohne Aufwand durchzuführen. In Bezug auf die Validität ist bspw. genau zu prüfen, ob ein gefundener Test auch wirklich genau das Personenmerkmal erfasst, das einen interessiert. Um das entscheiden zu können, benötigt man ein fundiertes wissenschaftliches Wissen über das zu testende Personenmerkmal. Allzu häufig lässt man sich jedoch von Alltagsbegriffen in die Irre führen. Diese sind jedoch meist unscharf definiert oder bedeuten oft etwas anderes als derselbe Begriff, wenn er im wissenschaftlichen Diskurs verwendet und definiert wird. Jede:r wird bspw. ein Gefühl dafür haben, was mit dem Begriff des Wissenschaftsverständnisses gemeint sein könnte. Vertieft man sich jedoch in die entsprechende wissenschaftliche Literatur, wird man feststellen, dass der Begriff nicht einheit-

lich definiert ist und je nach gewählter Definition unterschiedliche Facetten eines Wissenschaftsverständnisses betont werden. Mal mögen epistemologische Überzeugungen im Vordergrund stehen, mal Methodenkenntnisse, mal Kenntnisse innerhalb eines eng umgrenzten wissenschaftlichen Gebiets. Die jeweils gewählte wissenschaftlich zu begründende Definition muss sich letztlich auch in den Items eines Wissenschaftsverständnistests niederschlagen. Gute Tests definieren in einem Manual sehr genau das Personenmerkmal, das durch den Test erfasst wird. Diese Definition kann man dann mit der eigenen wissenschaftlich begründeten Definition abgleichen und so entscheiden, ob der Test für die eigenen Zwecke geeignet ist oder nicht.

Abgesehen von der klaren Definition des Personenmerkmals liefern gute Tests auch empirische Informationen über die Gütekriterien. Jeder gute Test liefert in seinem Manual Kennwerte zu Reliabilität, Objektivität und Validität. Sollten zu einem Test diese Informationen nicht vorliegen, ist man vermutlich gut beraten, diesen Test beiseite zu legen. Ähnlich verhält es sich mit den Anweisungen zur Durchführung, Auswertung und Interpretation eines Tests. Enthält ein Testmanual dazu nur Angaben, die viel Interpretationsspielraum lassen, sollte man von diesem Test wohl eher die Finger lassen.

Nicht in allen Fällen wird es möglich sein, ein gutes, etabliertes Testverfahren zu finden. In diesem Fall ist man auf Eigenkonstruktionen angewiesen. Die Ausführungen dieses Beitrags werden verdeutlicht haben, dass hierfür inhaltliche Expertise sowie methodische Kompetenzen im Bereich der Testentwicklung und -evaluation notwendig sind. Wenn letztere nicht zum prägenden Teil des eigenen Wissenschaftsgebiets zählen, sind Kooperationen mit den entsprechenden Disziplinen notwendig. Auch solche interdisziplinären Kooperationen erfordern (interdisziplinäre) Wissenschaftskommunikation. Diese kann manchmal vielleicht anstrengend sein, ist aber bereichernd für alle Beteiligten.

Literatur

Anderson LW, Krathwohl DR (Hrsg) (2001) A taxonomy for learning, teaching, and assessing: a revision of bloom's taxonomy of educational objectives. Longman, New York

Bromme R, Kienhues D (2014) Wissenschaftsverständnis und Wissenschaftskommunikation. In: Seidel T, Krapp A (Hrsg) Pädagogische Psychologie, 6. Aufl. Beltz, Weinheim, S 55–81

Cronbach LJ (1951) Coefficient alpha and the interval structure of tests. Psychometrika 16:297–334. https://doi.org/10.1007/BF02310555

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52:281–302. https://doi.org/10.1037/h0040957
de Ayala RJ (2009) The theory and practice of item response theory. Guilford, New York
Döring N, Bortz J (2016) Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler, 6. Aufl. Springer, Berlin
Embretson SE, Reise S (2000) Item response theory for psychologists. Erlbaum, Mahwah
Gulliksen H (1950) Theory of mental tests. Wiley, New York
Heckhausen H (1974) Leistung und Chancengleichheit. Hogrefe, Göttingen
Hartig J (2007) Skalierung und Definition von Kompetenzniveaus. In: Beck B, Klieme E (Hrsg) Sprachliche Kompetenzen. Konzepte und Messung – DESI-Studie. Beltz, Weinheim, S 72–82
Klauer KJ (1987) Kriteriumsorientierte Tests. Hogrefe, Göttingen
Kelava A, Moosbrugger H (2020) Einführung in die Item-Response-Theorie (IRT). In: Moosbrugger H, Kelava A (Hrsg), Testtheorie und Fragebogenkonstruktion, 3. Aufl. Springer, Berlin, S 369–409. https://doi.org/10.1007/978-3-662-61532-4_16
Klos S, Henke C, Kieren C, Walpuski M, Sumfleth E (2008) Naturwissenschaftliches Experimentieren und chemisches Fachwissen – zwei verschiedene Kompetenzen. Z Pädagog 54:304–321. https://doi.org/10.25656/01:4353
Land-Zandstra AM, Devilee JLA, Snik F, Buurmeijer F, van den Broek JM (2016) Citizen science on a smartphone: participants' motivations and learning. Public Underst Sci 25:45–60. https://doi.org/10.1177/0963662515602406
Masters K, Oh EY, Cox J, Simmons B, Lintott C, Graham G, Greenhill A, Holmes K (2016) Science learning via participation in online citizen science. J Sci Commun 15(03):A07. https://doi.org/10.22323/2.15030207
Reiss K, Sälzer C, Schiepe-Tiska A, Klieme E, Köller O (Hrsg) (2016) PISA 2015. Eine Studie zwischen Kontinuität und Innovation. Waxmann, Münster
Rost J (2004) Lehrbuch Testtheorie – Testkonstruktion, 2. Aufl. Huber, Bern
Schwinger M, Stiensmeier-Pelster J (2012) Erfassung von Self-Handicapping im Lern- und Leistungsbereich. Z Entwicklungspsychol Pädagog Psychol 44:68–80. https://doi.org/10.1026/0049-8637/a000061
van der Linden WJ (2016) Handbook of item response theory, volume one: models. Chapman & Hall, New York. https://doi.org/10.1201/9781315374512
Weinert FE (2001) Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: FE Weinert (Hrsg) Leistungsmessungen in Schulen. Weinheim, S 17–32
Weise G (1975) Psychologische Leistungstests. Hogrefe, Göttingen

Joachim Wirth ist Inhaber des Lehrstuhls für Lehr-Lernforschung der Ruhr-Universität Bochum. Er promovierte in der Psychologie an der Humboldt-Universität zu Berlin. Sein Forschungsschwerpunkt liegt im Bereich der Messung und Förderung selbstregulierten Lernens von Schüler:innen, Studierenden und Berufstätigen. In seiner experimentellen Feldforschung untersucht er den Einfluss instruktionaler Faktoren in formalen und nonformalen Lernkontexten.

Jens Fleischer ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Lehr-Lernforschung der Ruhr-Universität Bochum. Er promovierte in der Psychologie an der Universität Duisburg-Essen. Seine Forschungsschwerpunkte liegen im Bereich der Modellierung von Kompetenzen (Diagnostik und Assessment, Entwicklung und Validierung von Leistungs- und Kompetenztests), der fächerübergreifenden und fachbezogenen Problemlöseforschung sowie der Untersuchung kognitiver und nicht kognitiver Prädiktoren des Studienerfolgs.



Praxisbeitrag: Entwicklung und Überprüfung eines adaptierbaren Tests zum wissenschaftlichen Denken für Evaluationen in der Wissenschaftskommunikation

Till Bruckermann, Tanja M. Straka und Moritz Krell

Zusammenfassung

Wenn der Erfolg partizipativer Formate in der Wissenschaftskommunikation evaluiert werden soll, ist die Entwicklung von Fähigkeiten zum wissenschaftlichen Denken eine Zieldimension. Diese Fähigkeiten gehören neben dem Wissen zur naturwissenschaftlichen Grundbildung. Zur Evaluierung von Wissen und Fähigkeiten sind sorgfältig entwickelte Tests geeigneter als Fragebögen, da aus ihren Ergebnissen gültige Schlussfolgerungen über das tatsächliche Wissen oder die Fähigkeiten von Personen möglich sind. Im Praxisbeitrag wird eine Vorlage für einen Test über Fähigkeiten zum wissenschaftlichen Denken vorgestellt, der an unterschiedliche Fachkontexte adaptierbar ist, und gezeigt, inwiefern gültige Schlussfolgerungen aus diesem

T. Bruckermann (✉)
Institut für Erziehungswissenschaft, Leibniz Universität Hannover, Hannover, Deutschland
E-Mail: till.bruckermann@iew.uni-hannover.de

T. M. Straka
Institut für Ökologie, Technische Universität Berlin, Berlin, Deutschland
E-Mail: tanja.straka@tu-berlin.de

M. Krell
Didaktik der Biologie, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Deutschland
E-Mail: krell@leibniz-ipn.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_17

Test gezogen werden können. Dazu werden Fähigkeiten zum wissenschaftlichen Denken als psychologisches Konstrukt beschrieben und es wird dargestellt, wie diese Fähigkeiten bei Personen erfasst werden können und was bei der Entwicklung von Tests beachtet werden muss.

In öffentlichen Debatten über Erkenntnisse der Forschung können Fähigkeiten zum wissenschaftlichen Denken epistemisch fragwürdigen Überzeugungen – wie dem Glauben an pseudowissenschaftliche Erkenntnisse – entgegenwirken (Čavojová et al. 2019). Wenn Erkenntnisse aus unterschiedlichen Fachkontexten – wie beispielsweise zum Klimawandel oder zur Gentechnik – zur Debatte stehen, geht ein ausgeprägteres wissenschaftliches Denken mit individuellen Überzeugungen einher, welche eher wissenschaftlichem Konsens entsprechen (Drummond und Fischhoff 2017). Wissenschaftliches Denken gehört zur naturwissenschaftlichen Grundbildung (*scientific literacy;* Čavojová et al. 2019), die sowohl Wissen über Konzepte als auch über Prozesse der Wissensgenerierung in den Naturwissenschaften umfasst (Miller 2004). Im Folgenden soll geklärt werden, wie in der Wissenschaftskommunikation wissenschaftliches Denken von Personen mittels Tests erfasst werden kann.

In Tests zum wissenschaftlichen Denken sollen Fähigkeiten von Personen, wissenschaftliche Evidenz hinsichtlich ihrer Qualität einschätzen zu können (z. B. Drummond und Fischhoff 2017), abgebildet werden. In diesem Beitrag wird wissenschaftliches Denken im Sinne der Kognitionspsychologie als latentes Konstrukt aufgefasst (d. h. als nicht direkt beobachtbares Personenmerkmal; vgl. Bruckermann et al. 2021b). Dieses Konstrukt beschreibt das Problemlösen mit Denkweisen der Naturwissenschaften (Mathesius et al. 2019), wie dem Aufstellen von Hypothesen, Planen von Untersuchungen und Auswerten von Daten (Klahr und Dunbar 1988), sowie mithilfe naturwissenschaftlicher Arbeitsweisen, wie dem Beobachten, Experimentieren und Modellieren (Kind und Osborne 2017). Zumeist wird das Konstrukt wissenschaftliches Denken wie folgt getestet: Personen werden in Fragen mit mehreren Antwortoptionen (darunter eine korrekte; engl.: *single-choice questions*) darum gebeten, eine gültige Hypothese, einen gültigen Untersuchungsplan oder eine gültige Schlussfolgerung in einem Fachkontext auszuwählen (vgl. Opitz et al. 2017, für eine Übersicht). In der evaluativen Wissenschaftskommunikation (d. h., eine Untersuchung, inwieweit die Kommunikationsarbeit erfolgreich war) ist es allerdings aufwendiger und komplexer, das wissenschaftliche Denken mittels Tests zu erfassen, weil die Tests für unterschiedliche Fachkontexte (z. B. Biologie, Chemie, Physik) mit ihren spezifischen Themen entwickelt werden müssen. Für die Testentwicklung fehlen außerdem häufig die dafür nötigen Ressourcen.

Ziel dieses Praxisbeitrags ist es, eine Vorlage für einen Test vorzustellen, der an unterschiedliche Fachkontexte adaptierbar ist. Außerdem wird gezeigt, inwiefern aus diesem Test gültige Schlussfolgerungen über Fähigkeiten zum wissenschaftlichen Denken gezogen werden können. Dazu wird zunächst geklärt, inwiefern Fähigkeiten zum wissenschaftlichen Denken eine Zieldimension der Wissenschaftskommunikation sind, wie diese Fähigkeiten bei Personen erfasst werden können und was bei der Entwicklung von Tests beachtet werden muss, damit aus den Testergebnissen gültige Schlussfolgerungen über Personenfähigkeiten gezogen werden können.

1 Wissenschaftskommunikation zum wissenschaftlichen Denken

Wissenschaftliches Denken zu entwickeln und zu fördern (eine von mehreren Zieldimensionen der Wissenschaftskommunikation) gewinnt zunehmend an Bedeutung, wenn der Erfolg partizipativer Formate in der Wissenschaftskommunikation evaluiert werden soll (National Academies of Sciences, Engineering, and Medicine [NASEM] 2018). Durch den Paradigmenwechsel von *Scientific Literacy* zu *Public Engagement with Science* in der Wissenschaftskommunikation rückt nicht nur das Wissen über Wissenschaft, sondern auch die Beteiligung der Öffentlichkeit an Wissenschaft in den Vordergrund (vgl. Metag 2017, für eine Übersicht). Als Beispiel für partizipative Formate können Bürgerwissenschaftsprojekte angeführt werden, in denen sich die Teilnehmenden an der Produktion wissenschaftlichen Wissens beteiligen (Bonney et al. 2009) und darüber hinaus ihr Wissen, ihre Fähigkeiten und Einstellungen weiterentwickeln können (Phillips et al. 2018).

Neben Fähigkeiten zum wissenschaftlichen Arbeiten, wie beispielsweise zum Bestimmen von Tier- und Pflanzenarten (Crall et al. 2011), spielen Fähigkeiten zum wissenschaftlichen Denken eine zentrale Rolle in Bürgerwissenschaftsprojekten (NASEM 2018, Stylinski et al. 2020). Einerseits können Teilnehmende von Fähigkeiten zum wissenschaftlichen Denken profitieren, um sich umfassend an den Projektaktivitäten zu beteiligen (Burgess et al. 2017). Andererseits können diese Fähigkeiten ein Lernziel für Teilnehmende in Projekten sein (Phillips et al. 2018) und sie können auch andere Lernziele positiv beeinflussen (Edwards et al. 2017) – beispielsweise Verhaltensüberzeugungen, die ebenfalls eine Zieldimension der Wissenschaftskommunikation sind (Bruckermann et al. 2021a). Dennoch werden Fähigkeiten zum wissenschaftlichen Denken in Bürgerwissen-

schaftsprojekten seltener erhoben als beispielsweise Fähigkeiten zum Bestimmen von Tier- und Pflanzenarten (z. B. Crall et al. 2011; Stylinski et al. 2020).

In der evaluativen Wissenschaftskommunikation können für die Erhebung der individuellen Fähigkeiten zum wissenschaftlichen Denken unterschiedliche Verfahren eingesetzt werden, wie beispielsweise Fragebögen, Tests (z. B. Drummond und Fischhoff 2017) oder auch prozessorientierte Verfahren, wie eine direkte Verhaltensbeobachtung, bei der Teilnehmende ihre Fähigkeiten beispielsweise zum Datensammeln einsetzen müssen (z. B. Stylinski et al. 2020). In Bürgerwissenschaftsprojekten werden häufig Fragebögen verwendet, welche eine Selbsteinschätzung der eigenen Fähigkeiten zum wissenschaftlichen Denken erfassen oder Tests, welche die tatsächlichen Fähigkeiten überprüfen (vgl. Stylinski et al. 2020, für eine Übersicht). Wenn Teilnehmende in Fragebögen um eine Selbsteinschätzung ihrer Fähigkeiten auf einer Skala von beispielsweise 1 *(sehr gering)* bis 7 *(sehr hoch)* gebeten werden (sog. Selbstberichtsverfahren; siehe auch Wirth und Fleischer in diesem Band), nennt man die gewonnenen Daten auch Selbstberichte. Wenn Teilnehmende ihre Fähigkeiten zum wissenschaftlichen Denken nutzen müssen, um eine Aufgabe korrekt zu lösen, wie beispielsweise in Antwort-Wahlverfahren (z. B. *single-choice questions*), wird dies als Test bezeichnet. Aus den Ergebnissen eines Tests können – im Gegensatz zu Fragebögen – Schlussfolgerungen über das tatsächliche Wissen oder die Fähigkeiten zum wissenschaftlichen Denken der Teilnehmenden gezogen werden, vorausgesetzt, die Testergebnisse bilden die Fähigkeiten angemessen ab.

2 Gültigkeit von Tests in Bürgerwissenschaftsprojekten

Wenn wissenschaftliches Denken als eine Zieldimension von Wissenschaftskommunikation in die Evaluation des Projekterfolgs einbezogen wird (z. B. in Bürgerwissenschaftsprojekten), dann sollte der eingesetzte Test die Fähigkeiten zum wissenschaftlichen Denken von Teilnehmenden angemessen abbilden. Dazu müssen aus den Testergebnissen gültige Schlussfolgerungen gezogen werden können. Ob die gezogenen Schlussfolgerungen gültig sind, wird anhand des Testgütekriteriums der Validität beurteilt (siehe auch Böhmert und Abacioglu in diesem Band). Für Tests in Bürgerwissenschaftsprojekten sollte daher generell belegt werden, dass aus den Testergebnissen abgeleitete Schlussfolgerungen valide sind. Um die Validität der Schlussfolgerungen festzustellen, sollten zunächst unterschiedliche Aspekte der Validität geprüft und dann Belege aus der Prüfung angeführt werden (American Educational Research Association

[AERA] et al. 2014). Dazu werden im Folgenden exemplarisch drei Schritte unterschieden, die von der Theorie zum einsatzbereiten Test führen. Ein erster Schritt ist, das zu evaluierende Konstrukt klar zu definieren. Das heißt, die Merkmale des Konstrukts müssen mit Bezug zu Theorien und Modellen aus bisheriger Forschung beschrieben werden. In einem zweiten Schritt werden bei der Testkonstruktion gewisse Aufgabenmerkmale integriert, welche die aus der Theorie abgeleiteten Merkmale des Konstrukts widerspiegeln sollen. Außerdem sollte im dritten Schritt anhand der Aufgabenmerkmale gezeigt werden, dass die Fragen des Tests Denkprozesse initiieren, die auf das Konstrukt (hier: wissenschaftliches Denken) zurückzuführen sind. Weitere Belege für Validität, die hier nicht vertieft werden, können zum Beispiel durch Vergleiche mit anderen Konstrukten oder zwischen bewusst gewählten Stichproben gewonnen werden (AERA et al. 2014). Die drei beschriebenen Schritte werden im Folgenden auf das wissenschaftliche Denken angewandt und an einem Beispiel aus der Praxis verdeutlicht.

Im ersten Schritt wird das wissenschaftliche Denken als das zu evaluierende Konstrukt definiert. Dazu können in der Theorie mindestens zwei Sichtweisen identifiziert werden (NASEM 2018). Einerseits wird wissenschaftliches Denken als soziokulturelles Konstrukt aufgefasst, da die Prozesse der Wissensgenerierung und die Quellen des Wissens in Abhängigkeit von der jeweiligen Kultur beschrieben werden können. Andererseits wird es aus kognitionspsychologischer Sicht als die Fähigkeit verstanden, Probleme durch ein wissenschaftliches Vorgehen zu lösen, indem aufgestellte Vermutungen mit gewonnenen Belegen in Beziehung gesetzt werden (Mayer 2007). In diesem Problemlöseprozess werden oftmals drei Phasen unterschieden (Klahr und Dunbar 1988), die bestimmte Fähigkeiten erfordern und so wissenschaftliches Denken strukturieren. Den drei Phasen werden drei Teilfähigkeiten im wissenschaftlichen Denken zugeordnet: das Aufstellen von Hypothesen, das Planen von Untersuchungen und das Auswerten von Daten (z. B. Krell 2018). Da sich Tests auf die individuellen Fähigkeiten von Personen, also auf Personenmerkmale beziehen, wird im Folgenden wissenschaftliches Denken als Konstrukt aufgefasst.

Auf die Definition des wissenschaftlichen Denkens als Konstrukt folgt im zweiten Schritt die Aufgabenentwicklung. Die Definition des Konstrukts, das heißt, seine theoretisch bzw. in Modellen angenommene Struktur soll bestimmen, wie die Aufgaben des Tests konstruiert werden. Dazu zeigt die gesichtete Literatur, dass die zum wissenschaftlichen Denken angenommene Struktur unter anderem drei Teilfähigkeiten umfasst sowie weitere Denkprozesse, die sich auf das Erfassen der untersuchten Variablen sowie den Fachkontext beziehen (z. B. Krell 2018). Bei der Testkonstruktion sollte also berücksichtigt werden, welche

Denkprozesse zum Lösen der Aufgaben notwendig sind, da Schlussfolgerungen über individuelle Fähigkeiten in Tests auf den Testergebnissen beruhen – im Gegensatz zur direkten Beobachtung in prozessorientierten Verfahren (Shavelson 2013). Aus der angenommenen Struktur bzw. den Denkprozessen werden konkrete Aufgabenmerkmale abgeleitet, die in jeder Aufgabe des Tests enthalten sein sollten. In der hier vorgestellten, adaptierbaren Vorlage für einen Test werden drei Aufgabenmerkmale systematisch in den Test integriert, die das korrekte Lösen der Aufgaben durch wissenschaftliches Denken beeinflussen können, und zwar die durch die Aufgabe adressierte Teilfähigkeit, die Komplexität der in der Aufgabe beschriebenen Untersuchung und der Fachkontext der Untersuchung (Bruckermann et al. 2021b; Vorlage: Bruckermann et al., 2021c).

Tab. 1 verdeutlicht, dass alle Aufgaben im Test zum wissenschaftlichen Denken die drei zuvor genannten Aufgabenmerkmale berücksichtigen. Die Teilfähigkeiten zum wissenschaftlichen Denken sind in der adaptierbaren Vorlage als Aufgabenmerkmal integriert, indem sich die Aufgaben entweder mit dem Hypothesenaufstellen, dem Untersuchungenplanen oder dem Datenauswerten befassen. Ein Beispiel zum Datenauswerten ist die Aufgabe 12, welche in Abb. 1, links dargestellt ist. Die Komplexität einer Untersuchung ist als Aufgabenmerkmal in der Vorlage integriert, indem die Anzahl der unabhängigen Variablen, welche zum wissenschaftlichen Denken berücksichtigt werden müssen, variiert. In den Aufgaben werden drei unabhängige Variablen (A, B, C) unterschieden, von denen entweder nur eine unabhängige Variable (geringe Komplexität) oder zwei unabhängige Variablen variieren (hohe Komplexität; siehe Abb. 1, links). Die

Tab. 1 Übersicht der Aufgaben 1–18 im Test zum wissenschaftlichen Denken sowie die integrierten Aufgabenmerkmale Teilfähigkeiten, Komplexität und Fachkontext

	Komplexität (Anzahl der unabhängigen Variablen)	
	Niedrig = eine Variable	Hoch = zwei Variablen
Teilfähigkeiten		
Hypothesen aufstellen	Wildtierökologie (Aufgabe 1) Fledermausökologie (Aufgabe 7) Luftverschmutzung (Aufgabe 13)	Wildtierökologie (Aufgabe 4) Fledermausökologie (Aufgabe 10) Luftverschmutzung (Aufgabe 16)
Untersuchungen planen	Wildtierökologie (Aufgabe 5) Fledermausökologie (Aufgabe 11) Luftverschmutzung (Aufgabe 17)	Wildtierökologie (Aufgabe 2) Fledermausökologie (Aufgabe 8) Luftverschmutzung (Aufgabe 14)
Daten aus-werten	Wildtierökologie (Aufgabe 3) Fledermausökologie (Aufgabe 9) Luftverschmutzung (Aufgabe 15)	Wildtierökologie (Aufgabe 6) Fledermausökologie (Aufgabe 12) Luftverschmutzung (Aufgabe 18)

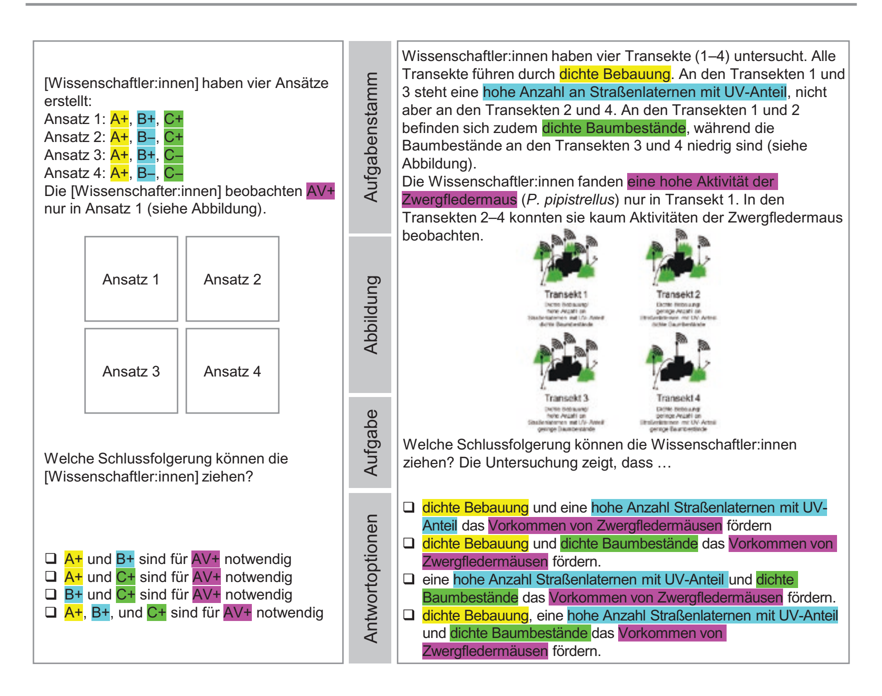


Abb. 1 Aufgabenbeispiel der adaptierbaren Vorlage bzw. des adaptierten Tests zum wissenschaftlichen Denken (Aufgabe 12). Anmerkung. Vorlage *(links)* und angepasste Testaufgabe *(rechts)* für den Fachkontext Fledermausökologie; A, B, C: unabhängige Variablen; AV: abhängige Variable; (+): Variable ausgeprägt; (−): Variable nicht ausgeprägt; farbige Hervorhebungen markieren korrespondierende Variablen aus Vorlage und Testaufgabe

unabhängigen Variablen (UV) und die abhängige Variable (AV) können entweder ausgeprägt (+) oder nicht ausgeprägt (−) sein. Darüber hinaus müssen die Fähigkeiten zum wissenschaftlichen Denken in verschiedenen Fachkontexten angewandt werden, sodass die Forschungsthemen in den Aufgaben systematisch variiert wurden (z. B. in der Fledermausökologie, siehe Abb. 1). Im folgenden Abschnitt wird beispielhaft erläutert, wie die Vorlage für drei Forschungsthemen im Rahmen der Evaluierung eines Bürgerwissenschaftsprojekts adaptiert wurde und dann im dritten Schritt belegt, dass die Aufgabenmerkmale tatsächlich die angenommenen Denkprozesse zum wissenschaftlichen Denken widerspiegeln.

3 Beispiel zur Testentwicklung in einem Bürgerwissenschaftsprojekt

Das vorgestellte Beispiel entstammt einem Forschungsprojekt, das drei Bürgerwissenschaftsprojekte umfasste, und zwar zur städtischen Wildtierökologie, zur städtischen Fledermausökologie und zur städtischen Luftverschmutzung. In allen drei Projekten konnten Teilnehmende in zeitlich begrenzten Feldphasen Daten erheben und auswerten. Obwohl sich die Projekte im Fachkontext ihrer Forschungsthemen unterschieden, verfolgten sie ähnliche Ziele, nämlich das Vorkommen von Tierarten bzw. von Luftverschmutzung im Zusammenhang mit weiteren Umweltvariablen im städtischen Kontext zu dokumentieren. Die kleinräumige Untersuchung des Vorkommens erfolgte im eigenen Garten oder auf abgesteckten Routen in Stadtteilen.

Nach einem Einführungsworkshop zum Forschungsvorhaben erhoben die Teilnehmenden während der Feldphasen Daten, indem sie eine Kamerafalle im Garten aufstellten (Bruckermann et al. 2021a), auf festgelegten Routen mittels Batlogger Fledermausrufe aufnahmen (Greving et al. 2022) oder auf individuellen Routen mittels Messrucksack Daten zur Luftverschmutzung erfassten (Tõnisson et al. 2021). Während der Feldphasen stand den Teilnehmenden eine Internetplattform zur Verfügung, die nicht nur Informationen zur Thematik, sondern auch ein Tutorial zur Bestimmung der erfassten Tierarten, Hilfsmittel zur Auswertung der erhobenen Daten im Zusammenhang mit weiteren Umweltvariablen sowie ein Forum zum persönlichen Austausch umfasste (siehe auch Bruckermann und Greving in diesem Band).

Durch den Test zum wissenschaftlichen Denken sollten die Forschungsfragen geklärt werden, inwiefern sich bei den Teilnehmenden die Fähigkeiten zum wissenschaftlichen Denken durch die Teilnahme an Bürgerwissenschaftsprojekten entwickeln und ob solche Fähigkeiten eher einen positiven Einfluss auf das Fachwissen über Ökologie oder Einstellungen zur Wissenschaft am Projektende haben (z. B. Bruckermann et al. 2021a). Um diese Forschungsfragen zu beantworten, wurde ein Test benötigt, der wissenschaftliches Denken in allen drei Fachkontexten der Bürgerwissenschaftsprojekte abbilden kann.

4 Adaptierung eines Tests zum wissenschaftlichen Denken

Der auf Grundlage einer bestehenden Vorlage (Bruckermann et al. 2021b; Krell 2018) adaptierte Test umfasste Forschungsthemen aus drei Fachkontexten (Wildtierökologie, Fledermausökologie und Luftverschmutzung), drei Teilfähigkeiten

des wissenschaftlichen Denkens (Hypothesen aufstellen, Untersuchungen planen, Daten auswerten) und zwei Komplexitätsstufen (eine unabhängige Variable und zwei unabhängige Variablen). Da der Test für drei Fachkontexte mit jeweils drei Teilfähigkeiten und zwei Komplexitätsstufen adaptiert werden musste, ergaben sich insgesamt $3 \times 3 \times 2 = 18$ Aufgaben (siehe Tab. 1). Um den Test anhand der Vorlage zu adaptieren und im jeweiligen Fachkontext einzubetten, wurden erstens Forschungsvorhaben im Fachkontext zu den jeweiligen Themen identifiziert (z. B. Wirkung von nächtlichem Kunstlicht und Baumbewuchs auf Fledermäuse: Straka et al. 2019). Zweitens wurden die Angaben der Forschungsvorhaben zu den unabhängigen und abhängigen Variablen, Hypothesen, dem Untersuchungsplan und den Schlussfolgerungen aus den Daten entnommen. Drittens wurden Angaben der ausgewählten Forschungsvorhaben zur Gestaltung der Testaufgaben für eine der drei Teilfähigkeiten (d. h. Hypothesen aufstellen, Untersuchungen planen und Daten auswerten) sowie eine der zwei Komplexitätsstufen der Untersuchungen übernommen (d. h. eine oder zwei unabhängige Variablen variieren). Die Anpassung des Tests anhand der Vorlage wird beispielhaft für die Aufgabe 12 vorgestellt (vollständiger Test: Bruckermann et al. 2021c).

In dem für die vorgestellten Bürgerwissenschaftsprojekte adaptierten Test betraf Aufgabe 12 die Teilfähigkeit der Datenauswertung, die Anzahl der unabhängigen Variablen war zwei (d. h. hohe Komplexität der Untersuchung), und der Fachkontext war die Fledermausökologie (vgl. Abb. 1, rechts). Im Aufgabenstamm wird den Teilnehmenden zunächst der Untersuchungsplan für das untersuchte Phänomen (hier: das Vorkommen von Fledermäusen und deren Aktivität als abhängige Variable) vorgestellt. Für jede variierte, unabhängige Variable (hier: Anzahl von Straßenlaternen, Baumbestand) wird beschrieben, ob sie in den vier Ansätzen (hier: Transekte = festgelegte Wegstrecken) ausgeprägt (+) oder nicht ausgeprägt (−) war. Im Beispiel der Aufgabe 12 lagen die vier Transekte in dicht bebautem Gebiet mit hoher (+) oder niedriger (−) Anzahl von Straßenlaternen sowie dichtem (+) oder weniger dichtem (−) Baumbestand (vgl. Abb. 1, rechts). Anschließend wurde die von den Wissenschaftler:innen gemachte Beobachtung, dass Zwergfledermäuse (*Pipistrellus pipistrellus*) auf dem Transekt 1 aktiver waren als auf den übrigen Transekten, dargestellt. In einer Abbildung wurden die vier Ansätze nochmals zusammengefasst. Die Aufgabe forderte die Teilnehmenden zur Entscheidung auf, welche Schlussfolgerung aus dieser Beobachtung gezogen werden kann. In den Antwortoptionen wurden vier mögliche Schlussfolgerungen vorgegeben. Die Schlussfolgerungen zu der Beobachtung basierten auf verschiedenen Kombinationen der unabhängigen Variablen. Nur eine Kombination der unabhängigen Variablen in den Antwortoptionen beschreibt eine zulässige Schlussfolgerung. Die Teilnehmenden sollten diese Kombination identifizieren und die entsprechende Antwortoption ankreuzen.

5 Überprüfung des Tests im Feld

Das Ziel der hier beschriebenen Überprüfung des Tests im Feld war, ob aus den Testergebnissen des adaptierten Tests gültige Schlussfolgerungen auf die Fähigkeiten der Teilnehmenden zum wissenschaftlichen Denken gezogen werden können. Um die Gültigkeit der Schlussfolgerungen zu belegen, wurde überprüft ob die in der Konstruktion berücksichtigten Aufgabenmerkmale beeinflussen, wie schwer die Aufgaben zu lösen sind. Dabei sollte ein ausreichend großer Teil der Schwierigkeit des gesamten Tests auf die Aufgabenschwierigkeit der kombinierten Aufgabenmerkmale zurückzuführen sein (z. B. $R^2 > 25\,\%$; Hartig und Frey 2012). Die Überprüfung im Feld erfolgte mit einer Stichprobe von 374 Teilnehmenden des Bürgerwissenschaftsprojekts Wildtierforscher in Berlin (Wildtierökologie), die im Mittel ca. 53 Jahre alt waren und überwiegend höhere Bildungsabschlüsse hatten. Die Teilnehmenden füllten alle 18 Aufgaben des Tests vor ihrer Projektteilnahme aus, um eine Verzerrung der Testergebnisse durch Lerneffekte während des Projekts zu vermeiden (Bruckermann et al. 2021b).

Die Daten der Teilnehmenden aus dem Test wurden mittels zweier Verfahren der *Item-Response-Theory* analysiert (siehe auch Wirth und Fleischer in diesem Band). Wie die Verfahren für die im Folgenden beschriebenen Analyse angewandt wurden, sollte nachgelesen werden (z. B. Bruckermann et al. 2021b), weil die Verfahren eine gewisse methodische Expertise erfordern. Das erste Verfahren beschreibt einen Personenfähigkeitsparameter (θ_s; Ausprägung des Personenmerkmals) sowie einen Aufgabenschwierigkeitsparameter (β_i; Schwierigkeit einer Aufgabe) durch ein Modell, das wissenschaftliches Denken holistisch als eindimensionales Konstrukt betrachtet (sogenanntes Einparametrisch-Logistisches Modell: 1PLM; vgl. Krell 2018). Dieses Modell wies eine ausreichende Passung zu den Daten auf und zeigte, dass sowohl die Aufgabenschwierigkeiten hinreichend unterschiedlich waren und mit den Aufgaben auch die Teilnehmenden im Hinblick auf ihre Fähigkeiten unterschieden werden konnten (Bruckermann et al. 2021b). Um prüfen zu können, ob die in der Vorlage angenommenen und im Test adaptierten Aufgabenmerkmale zur Schwierigkeit der Aufgaben beitragen, wurden die Daten mit einem weiteren Verfahren im Rahmen der *Item-Response-Theory* analysiert. In diesem Verfahren wird ein Modell angelegt, in dem der Aufgabenschwierigkeitsparameter (β'_i) aus einer Kombination der den Aufgabenmerkmalen zugeordneten Schwierigkeitsparametern (α_k) gebildet wird (sogenanntes Linear Logistisches Test-Modell: LLTM: vgl. Krell 2018). Die Schwierigkeitsparameter der Aufgabenmerkmale (α_k) waren alle von null verschieden, das heißt, sie trugen signifikant zur Aufgabenschwierigkeit (β'_i) bei.

Beispielsweise waren Testaufgaben zum Datenauswerten (Aufgabenmerkmal Teilfähigkeit) deutlich schwieriger als Aufgaben zur Untersuchungsplanung und auch die Fachkontexte unterschieden sich in der Schwierigkeit (Aufgabenmerkmal Fachkontext). Die Aufgabenschwierigkeitsparameter des ersten Modells (β_i) und des zweiten Modells (β'_i) korrelierten stark positiv miteinander.

Zusammengefasst belegen die Analysen, dass die Teilfähigkeiten, die Komplexität der Untersuchung und die Fachkontexte als Aufgabenmerkmale wie angenommen die Schwierigkeit des Tests zum wissenschaftlichen Denken beeinflussen. Bei der Adaptierung des Tests für andere Fachkontexten in der evaluativen Wissenschaftskommunikation sollten diese Merkmale berücksichtigt werden, um gültige Schlussfolgerungen über die Fähigkeiten von Teilnehmenden zum wissenschaftlichen Denken ziehen zu können.

6 Fazit und Ausblick

In der evaluativen Wissenschaftskommunikationsforschung sollte sichergestellt werden, dass die aus Testergebnissen gewonnenen Schlussfolgerungen über Personenmerkmale wie Wissen und eben auch Fähigkeiten – beispielsweise zum wissenschaftlichen Denken – gültig sind. Die Gültigkeit sollte im Feld für die jeweilige Stichprobe überprüft werden. Die Testentwicklung bzw. -adaptierung sollte auf etablierte Theorien und Modelle zu dem jeweiligen Konstrukt gestützt werden und möglichst auf bereits etablierte Tests zurückgreifen (z. B. Drummond und Fischhoff 2017), die im besten Fall eine Vorlage zur Adaptierung, wie hier eines Tests zum wissenschaftlichen Denken bieten (Bruckermann et al. 2021b). Wenn kein Test zur Verfügung steht, weil beispielsweise themenspezifisches Fachwissen erhoben werden soll (vgl. Bruckermann et al. 2022), dann sollten theoretisch angenommene Aufgabenmerkmale die Testkonstruktion leiten.

Die sorgfältige Entwicklung eines Tests setzt voraus, dass das zu erfassende Personenmerkmal bekannt und klar definiert ist. Häufig können aber in der Evaluation partizipativer Formate wie beispielsweise in Bürgerwissenschaftsprojekten nicht alle zu erfassenden Konstrukte antizipiert werden, da es sich um informelles und nicht-geplantes Lernen handelt, oder die Anzahl der Konstrukte ist so umfangreich, dass die Erfassung zeitlich aufwendig ist (Phillips et al. 2018). Um den mit Tests verbundenen Zeitaufwand für Teilnehmende zu vermeiden und einen größeren Umfang an Konstrukten abzudecken, wird oftmals auf Fragebögen zurückgegriffen, die allerdings nur selbstberichtete Einschätzungen zu den Konstrukten enthalten (z. B. Peter et al. 2021). Im Gegensatz zu Fragebögen können aus den Ergebnissen sorgfältig entwickelter Tests gültige

Schlussfolgerungen über das tatsächliche Wissen oder die Fähigkeiten der Teilnehmenden gezogen werden (siehe auch Wirth und Fleischer in diesem Band). Gültige Schlussfolgerungen sind insbesondere dann möglich, wenn schon bei der Testkonstruktion die Aufgaben derart gestaltet werden, dass sie das erfasste Konstrukt widerspiegeln, und außerdem nachgewiesen werden kann, dass diese Aufgabenmerkmale die Schwierigkeit des Tests wie angenommen beeinflussen.

Beim Einsatz bereits publizierter Tests sollte beachtet werden, dass die Gültigkeit von Schlussfolgerungen aus den Testergebnissen stets in einer bestimmten Stichprobe nachgewiesen wurde, wie hier bei Bürgerwissenschaftler:innen mit einem höheren Bildungsgrad. Publizierte Tests sollten also in Stichproben eingesetzt werden, in denen die Befragten ähnliche Personenmerkmale aufweisen, und die Gültigkeit von Schlussfolgerungen aus den Testergebnissen sollte bestenfalls erneut geprüft werden. Dazu kann zunächst eine Pilotstudie mit kleinerer Stichprobe durchgeführt werden, die aber eine ähnliche Personengruppe wie die später befragte umfasst. Außerdem sollten publizierte Tests nur dann verändert werden, indem beispielsweise Aufgaben umformuliert oder ausgelassen werden, wenn die Veränderung begründet werden kann. Außerdem kann auf adaptierbare Vorlagen zurückgegriffen werden, welche die Anpassung des Tests für die Stichprobe vorsehen. Nach der Veränderung oder Anpassung eines Tests sollte die Gültigkeit ebenfalls überprüft werden.

Tests sollten zu mindestens zwei Zeitpunkten eingesetzt werden, um in längsschnittlichen Studien Rückschlüsse auf die Entwicklung von Wissen oder Fähigkeiten ziehen zu können (siehe auch Böhmert und Abacioglu in diesem Band). So konnte eine längsschnittliche Studie in einem Bürgerwissenschaftsprojekt zeigen, dass Fähigkeiten der Teilnehmenden zum wissenschaftlichen Denken einen positiven Einfluss auf ihre Verhaltensüberzeugungen haben (Bruckermann et al. 2021a). Dazu wurden die Ergebnisse eines Tests zum wissenschaftlichen Denken und die Daten eines Fragebogens zu Verhaltensüberzeugungen aus Befragungen vor und nach einem Bürgerwissenschaftsprojekts auf Zusammenhänge untersucht. Des Weiteren sind Tests in Experimenten notwendig um Veränderungen in bestimmten Konstrukten wie beispielsweise Wissen auf die untersuchten Faktoren zurückführen zu können (siehe auch Stadtler und Schuster in diesem Band).

Literatur

American Educational Research Association [AERA], American Psychological Association, National Council on Measurement in Education (2014) Standards for educational and psychological testing. American Educational Research Association, Washington, DC

Bonney RE, Cooper CB, Dickinson J, Kelling S, Phillips TB, Rosenberg KV, Shirk J (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59:977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bruckermann T, Greving H, Schumann A, Stillfried M, Börner K, Kimmig SE, Hagen R, Brandt M, Harms U (2021a) To know about science is to love it? Unraveling cause–effect relationships between knowledge and attitudes toward science in citizen science on urban wildlife ecology. J Res Sci Teach 58:1179–1202. https://doi.org/10.1002/tea.21697

Bruckermann T, Straka TM, Stillfried M, Krell M (2021b) Context matters: accounting for item features in the assessment of citizen scientists' scientific reasoning skills. Citizen Sci Theor Pract 6. https://doi.org/10.5334/cstp.309

Bruckermann T, Straka TM, Stillfried M, Krell M (2021c) Context matters: accounting for item features in the assessment of citizen scientists' scientific reasoning skills [Supplemental material]. Citizen Sci Theor Pract 6. https://doi.org/10.5334/cstp.309.s1

Bruckermann T, Stillfried M, Straka TM, Harms U (2022) Citizen science projects require agreement: a Delphi study to identify which knowledge on urban ecology is considered relevant from scientists' and citizens' perspectives. Int J Sci Educ Part B Commun Public Engagem 12:75–92. https://doi.org/10.1080/21548455.2022.2028925

Burgess HK, DeBey LB, Froehlich HE, Schmidt N, Theobald EJ, Ettinger AK, Hille Ris Lambers J, Tewksbury J, Parrish JK (2017) The science of citizen science: exploring barriers to use as a primary research tool. Biol Cons 208:113–120. https://doi.org/10.1016/j.biocon.2016.05.014

Čavojová V, Šrol J, Jurkovič M (2019) Why should we try to think like scientists? Scientific reasoning and susceptibility to epistemically suspect beliefs and cognitive biases. Appl Cogn Psychol 34:85–95. https://doi.org/10.1002/acp.3595

Crall AW, Newman GJ, Stohlgren TJ, Holfelder KA, Graham J, Waller DM (2011) Assessing citizen science data quality: an invasive species case study. Conserv Lett 4:433–442. https://doi.org/10.1111/j.1755-263X.2011.00196.x

Drummond C, Fischhoff B (2017) Development and validation of the scientific reasoning scale. J Behav Decis Mak 30:26–38. https://doi.org/10.1002/bdm.1906

Edwards R, McDonnell D, Simpson I, Wilson A (2017) Educational backgrounds, project design, and inquiry learning in citizen science. In: Herodotou C, Sharples M, Scanlon E (Hrsg) Citizen inquiry. Synthesising science and inquiry learning. Routledge, Abingdon, Oxon, New York, NY, S 195–209

Greving H, Bruckermann T, Schumann A, Straka TM, Lewanzik D, Voigt-Heucke S, Marggraf L, Lorenz J, Brandt M, Voigt CC, Harms U, Kimmerle J (2022) Improving attitudes and knowledge in a citizen science project about urban bat ecology. Ecol Soc 27. https://doi.org/10.5751/ES-13272-270224

Hartig J, Frey A (2012) Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. Psychol Rundsch 63:43–49. https://doi.org/10.1026/0033-3042/a000109

Kind P, Osborne J (2017) Styles of scientific reasoning: a cultural rationale for science education? Sci Educ 101:8–31. https://doi.org/10.1002/sce.21251

Klahr D, Dunbar K (1988) Dual space search during scientific reasoning. Cogn Sci 12:1–48. https://doi.org/10.1207/s15516709cog1201_1

Krell M (2018) Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choie-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie. Z Didakt Nat 24:1–15. https://doi.org/10.1007/s40573-017-0069-0
Mathesius S, Krell M, Upmeier zu Belzen A, Krüger D (2019) Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables. Z Pädagog 4:492–510. https://doi.org/10.25656/01:23991
Mayer J (2007) Erkenntnisgewinnung als wissenschaftliches Problemlösen. In: Krüger D, Vogt H (Hrsg) Theorien in der biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden. Springer, Berlin, S 177–186
Metag J (2017) Rezeption und Wirkung öffentlicher Wissenschaftskommunikation. In: Bonfadelli H, Fähnrich B, Lüthje C, Milde J, Rhomberg M, Schäfer MS (Hrsg) Forschungsfeld Wissenschaftskommunikation. Springer Fachmedien Wiesbaden, Wiesbaden, S 251–274
Miller JD (2004) Public understanding of, and attitudes toward, scientific research: what we know and what we need to know. Public Underst Sci 13:273–294. https://doi.org/10.1177/0963662504044908
National Academies of Sciences, Engineering, and Medicine [NASEM] (2018) Learning through citizen science: Enhancing opportunities by design. National Academies Press (US), Washington (DC)
Opitz A, Heene M, Fischer F (2017) Measuring scientific reasoning – a review of test instruments. Educ Res Eval 23:78–101. https://doi.org/10.1080/13803611.2017.1338586
Peter M, Diekötter T, Kremer K, Höffler T (2021) Citizen science project characteristics: connection to participants' gains in knowledge and skills. PLoS ONE 16:e0253692. https://doi.org/10.1371/journal.pone.0253692
Phillips TB, Porticella N, Constas M, Bonney RE (2018) A framework for articulating and measuring individual learning outcomes from participation in citizen science. Citizen Sci Theor Pract 3:3. https://doi.org/10.5334/cstp.126
Shavelson RJ (2013) On an approach to testing and modeling competence. Educ Psychol 48:73–86. https://doi.org/10.1080/00461520.2013.779483
Straka TM, Wolf M, Gras P, Buchholz S, Voigt CC (2019) Tree cover mediates the effect of artificial light on urban bats. Front Ecol Evol 7:27. https://doi.org/10.3389/fevo.2019.00091
Stylinski CD, Peterman K, Phillips TB, Linhart J, Becker-Klein R (2020) Assessing science inquiry skills of citizen science volunteers: a snapshot of the field. Int J Sci Educ Part B Commun Public Engagem 10:77–92. https://doi.org/10.1080/21548455.2020.1719288
Tõnisson L, Voigtländer J, Weger M, Assmann D, Käthner R, Heinold B, Macke A (2021) Knowledge transfer with citizen science: luft-leipzig case study. Sustainability 13:7855. https://doi.org/10.3390/su13147855

Till Bruckermann ist Universitätsprofessor an der Leibniz Universität Hannover. Er forscht zu informellem Lernen in Bürgerwissenschaftsprojekten und insbesondere zur Entwicklung eines Wissenschaftsverständnisses.

Tanja M. Straka ist wissenschaftliche Mitarbeiterin an der Technischen Universität Berlin. Sie forscht zu urbaner Biodiversität und Mensch-Tier-Umwelt-Beziehungen im urbanen Raum, hierunter auch im Zusammenhang mit Bürgerwissenschaftsprojekten.

Moritz Krell ist stellvertretender Direktor der Abteilung Didaktik der Biologie am IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik und Professor für Didaktik der Biologie an der Christian-Albrechts-Universität zu Kiel. Einer seiner Forschungsschwerpunkte ist die Erfassung und Förderung von Kompetenzen wissenschaftlichen Denkens.



Experimentelle Herangehensweisen in der Evaluation von Maßnahmen der Wissenschaftskommunikation

Marc Stadtler und Corinna Schuster

Zusammenfassung

In unserem Beitrag zeigen wir Potenziale und Limitationen von Experimenten in der Evaluation von Maßnahmen der Wissenschaftskommunikation auf. Dabei arbeiten wir zunächst den Unterschied zwischen experimentellen und nicht-experimentellen Studien heraus und zeigen deren spezifische Erkenntnismöglichkeiten auf. Die Logik von Experimenten wird anhand zahlreicher Beispiele aus der Literatur verdeutlicht. Das Kapitel schließt mit einem Blick auf methodische Herausforderungen und deren Bewältigung, die beim Experimentieren in der angewandten Forschung der Wissenschaftskommunikation besonders verbreitet sind.

1 Einführung

Lässt sich mit authentischen Ausstellungsstücken die Aufmerksamkeit von Museumsbesucher:innen besonders gut gewinnen (vgl. Schwan et al. 2016)? Führen Citizen-Science-Projekte zu einem verbesserten Wissenschaftsverständnis bei den teilnehmenden Bürger:innen (vgl. Bonney et al. 2015)? Fördert die Dar-

M. Stadtler (✉) · C. Schuster
Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Bochum, Deutschland
E-Mail: marc.stadtler@rub.de

C. Schuster
E-Mail: corinna.schuster@rub.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_18

stellung wissenschaftlicher Sachverhalte in Comics das Themenverständnis bei jungen Menschen (vgl. Lin et al. 2015)?

Studien, in denen Maßnahmen der Wissenschaftskommunikation evaluiert werden, rücken oft die Frage nach der Wirkung spezifischer kommunikativer Maßnahmen in den Mittelpunkt. Wünschenswerte Effekte in Bereichen wie Interessensförderung, Vertrauensbildung oder Wissenschaftsverständnis sollen in einem kausalen Sinne auf die zu evaluierende Maßnahme zurückgeführt werden. Mit diesem Erkenntnisinteresse kommen Studiendesigns, die Reaktionen von Bürger:innen, Museumbesucher:innen oder Podcasthörer:innen ausschließlich beobachten, ohne in den Rezeptionsprozess einzugreifen, an ihre Grenzen. Stattdessen sind Experimente die Methode der Wahl, die ihren Ursprung in den Naturwissenschaften haben und im Fokus dieses Beitrags stehen. Der Beitrag gibt eine praxisnahe Einführung in die Logik des Experiments, skizziert seine verschiedenen Erscheinungsformen und zeigt Potenziale und Limitationen des Experimentierens im Kontext der Wissenschaftskommunikation auf.

2 Grundbegriffe des Experimentierens

Stellen Sie sich vor, ein Team von Forscher:innen möchte herausfinden, ob der Besuch von Science Slams zu einer positiven Einstellung gegenüber Wissenschaft führt. Die Forschenden besuchen eine ganze Reihe solcher Veranstaltungen, erheben von den anwesenden Besucher:innen Daten zur Häufigkeit des Besuchs von Science Slams und messen ihre Einstellung zur Wissenschaft. Zur Freude des Forscher:innenteams zeigen die statistischen Analysen einen positiven Zusammenhang der Variablen: Je öfter Personen Science Slams besuchen, desto positiver ihre Einstellung gegenüber der Wissenschaft. Lässt sich hieraus ableiten, dass Unterschiede in den Einstellungen kausal auf die Häufigkeit des Besuchs von Science Slams zurückzuführen sind? Tatsächlich ist die im fiktiven Beispiel beobachtete Kontiguität zwischen Ursache und Wirkung, also ihr gemeinsames Auftreten, eines von mehreren Kriterien, die erfüllt sein müssen, um auf einen kausalen Zusammenhang zwischen zwei Variablen schließen zu können (für einen Überblick der erkenntnistheoretischen Grundlagen von Kausalschlüssen, siehe Field und Hole 2003). Allerdings ist die Kontiguität nur ein notwendiges, kein hinreichendes Kriterium. Um auf eine kausale Beziehung zwischen zwei Variablen zu schließen, muss die Ursache zudem zeitlich *vor* der Wirkung liegen. Dieses Kriterium zeigt ein erstes Problem der oben skizzierten querschnittlichen Studienanlage auf, denn es wäre auch denkbar, dass die positiven Wissenschaftseinstellungen der Science-Slam-Enthusiast:innen bereits vor dem Besuch der

Science Slams vorlagen, mithin nicht als Wirkung infrage kommen. Schließlich könnte es auch eine dritte, von Bauer und Kolleg:innen (2014) als „vergessene Variable" bezeichnete Größe geben, die *sowohl* Wissenschaftseinstellung *als auch* Häufigkeit der Besuche von Science Slams beeinflusst. So könnte die Nähe des von den Versuchspersonen ausgeübten Berufs zum Wissenschaftsbetrieb sowohl einen Einfluss darauf ausüben, wie häufig man Science Slams besucht, als auch darauf, wie positiv man der Wissenschaft gegenübersteht. Eine solche Konfundierung kann nur ausgeschlossen werden, wenn ein Studiendesign gewählt wird, in dem die Forschenden die mutmaßlich verursachende Variable isolieren und sie in einer Untersuchungsbedingung präsentieren, während diese in einer anderen Bedingung, die ansonsten vergleichbar ist, nicht dargeboten wird. Das Ausmaß der Differenz zwischen den Bedingungen, in denen die Ursache präsent beziehungsweise nicht präsent ist, nährt laut Field und Hole das Vertrauen darin, dass es sich um einen kausalen Zusammenhang zweier Variablen handelt.

Genau das ist die Grundidee des wissenschaftlichen Experimentierens: Veränderungen in der Ausprägung einer *abhängigen Variable* werden durch gezielte Manipulationen einer anderen Variable, die als *unabhängige Variable* bezeichnet wird, verursacht. Anders als im weiter oben dargestellten korrelativen Design, greift der:die Forscher:in aktiv in den Rezeptionsprozess ein und manipuliert eine (oder mehrere) Variable(n), während er:sie andere kontrolliert. Durch das hohe Maß an Kontrolle steigt die eindeutige Interpretierbarkeit der Ergebnisse (man spricht von einer hohen ‚internen Validität'), weil Veränderungen der abhängigen Variable mit nur einer Ursache plausibel erklärt werden können. Das Gegenstück zur internen Validität ist die externe Validität, also die Verallgemeinerbarkeit der Ergebnisse eines Experiments über die Untersuchungssituation hinaus. Hier zeigen Experimente mitunter Defizite, weil ein hohes Maß an experimenteller Kontrolle zu artifiziellen Untersuchungssituationen führen kann, die sich von der natürlichen Auseinandersetzung der Bürger:innen mit Wissenschaftsinhalten deutlich unterscheiden. Ein konstruktiver Umgang mit diesem Problem könnte darin bestehen, zunächst sorgfältig im Labor getestete Wirkhypothesen zunehmend auch in authentischeren Settings zu überprüfen, um das Beste aus beiden Welten zu vereinen.

3 Experimentaldesigns für die Evaluation von Maßnahmen der Wissenschaftskommunikation

Wer eine Maßnahme der Wissenschaftskommunikation mithilfe eines Experiments evaluieren will, muss im Rahmen der Studienplanung Entscheidungen über das Experimentaldesign treffen. Hiermit bezeichnet man die Auswahl von

abhängigen und unabhängigen Variablen sowie die Anordnung der Messzeitpunkte innerhalb einer Studie. Die Wahl eines Designs will wohlüberlegt sein, hat sie doch gewichtige Konsequenzen für die mit dem Experiment verbundenen Erkenntnismöglichkeiten, den benötigten Stichprobenumfang und – hierüber vermittelt – den mit der Studiendurchführung verbundenen Aufwand.

Kennzeichnend für sogenannte Messwiederholungsdesigns (engl.: repeated measures design) ist, dass jede:r Untersuchungsteilnehmende alle Experimentalbedingungen durchläuft. Ein solches Design wurde zum Beispiel von Scharrer und Kolleg:innen (2017) in ihrer Laborstudie zum Einfachheitseffekt der Wissenschaftspopularisierung (Scharrer et al. 2017) verwendet. In dem Experiment lasen medizinische Lai:innen authentische Wissenschaftsartikel aus dem Internet, in denen medizinische Themen erläutert wurden. Die Artikel richteten sich entweder an ein Expert:innenpublikum (z. B. auf www.aerzteblatt.de) und waren sprachlich komplex oder an ein Lai:innenpublikum (z. B. auf www.manshealth.com) und verwendeten eine stark vereinfachte Sprache. Weil die Proband:innen nacheinander Artikel (zu verschiedenen Themen) aus beiden Kategorien lasen, handelt es sich um ein Messwiederholungsdesign (hier mit dem Messwiederholungsfaktor „Sprachliche Einfachheit"). Scharrer und Kolleg:innen (2017) vermuteten unter anderem, dass die Untersuchungsteilnehmer:innen Behauptungen aus den sprachlich einfacheren Lai:innentexten mehr Glauben schenken als den Behauptungen aus den sprachlich komplexeren Artikeln, die sich an ein Expert:innenenpublikum richteten. Die Ergebnisse bestätigten die Hypothese der Studienautor:innen.

Weil die mit statistischen Verfahren zu vergleichenden Mittelwerte aus den Angaben derselben Versuchspersonen gebildet werden, benötigen Messwiederholungsdesigns vergleichsweise wenige Proband:innen und sind damit ökonomischer als Zwischensubjektdesigns, in denen die Mittelwerte basierend auf den Angaben *verschiedener* Versuchspersonen gebildet werden. Zu den Nachteilen von Messwiederholungsdesigns zählt jedoch, dass Proband:innen die mit der experimentellen Manipulation verbundene Intention erkennen und sich zum Beispiel möglichst kooperativ im Sinne der vermuteten Hypothese verhalten könnten (auch das Gegenteil ist denkbar), was einem verzerrten Antwortverhalten gleichkäme. Zudem besteht die Gefahr, dass Unterschiede zwischen den Experimentalbedingungen im Sinne von Kontrasteffekten überzeichnet und Effekte der experimentellen Manipulation überschätzt werden. So ist es in der Studie von Scharrer und Kolleg:innen denkbar, dass die einfachen Themendarstellungen besonders einfach wirken, wenn sie direkt nach einer komplexen Darstellung rezipiert werden und umgekehrt. Anders ausgedrückt: Die Messungen können potenziell durch Erfahrungen beeinflusst werden, die Proband:innen in

anderen Versuchsbedingungen gemacht haben. Solche Phänomene, die man auch als Spill-Over-Effekte bezeichnet, gilt es zu vermeiden, weil sie die Interpretierbarkeit der Studienergebnisse einschränken.

Zwischensubjektdesigns begegnen diesem Problem, indem den Stufen einer unabhängigen Variable distinkte Subgruppen der Gesamtstichprobe zugeordnet werden. So wurden zum Beispiel in einer Studie von Lin und Kolleg:innen (2015) erwachsenen Proband:innen Informationen zum Thema Nanotechnologie dargeboten. Die Hälfte der Proband:innen erhielt die Information in Form eines Sachtexts, während die andere Hälfte der Proband:innen dieselben Informationen als Comic aufbereitet rezipierten. Es zeigte sich, dass die Teilnehmer:innen der Comic-Bedingung mehr Freude beim Lernen erlebten und ein größeres Interesse an Wissenschaft bekundeten, als die Proband:innen der Sachtextbedingung, wohingegen keine Unterschiede in den Bereichen Wissenserwerb und Einstellungen zur Wissenschaft gefunden wurden.

Der Vorteil von Zwischensubjektdesigns besteht darin, dass die im Zusammenhang mit Innersubjektdesigns diskutierten Nachteile (Hypothesenraten und Kontrasteffekte) keine bzw. eine untergeordnete Rolle spielen. Sie sind durch die größeren Stichprobenumfänge allerdings aufwendiger in der Durchführung. Zudem gilt, dass selbst bei randomisierter Zuweisung der Versuchspersonen zu Experimentalbedingungen das oben bereits diskutierte Drittvariablenproblem nicht vollständig ausgeschlossen werden kann. Hier gilt es bereits im Vorfeld, potenziell konfundierende Variablen wie Wissenschaftsinteresse oder thematisches Vorwissen zu identifizieren und diese Konstrukte als Kontrollvariablen zu erfassen. Dies gibt den Forschenden zumindest die Möglichkeit, den Effekt der konfundierenden Variablen statistisch mit dem Verfahren der Kovarianzanalyse zu kontrollieren (für eine detaillierte Darstellung dieses Verfahrens, siehe Döring und Bortz 2016).

In der Evaluation von Maßnahmen der Wissenschaftskommunikation sind neben den beiden vorgestellten Grundformen vor allem gemischte Designs mit Zwischensubjekt- und Innersubjektfaktor angezeigt. Dabei werden verschiedene Varianten einer Maßnahme als Stufen eines Zwischensubjektfaktors miteinander verglichen. Die interessierenden Variablen werden zu mindestens zwei Messzeitpunkten erhoben. In der bereits vorgestellten Studie von Lin und Kolleg:innen (2015) wurde das Wissen über Nanotechnologie, die Einstellungen zur Nanotechnologie und das Interesse am Thema zum Beispiel unmittelbar vor dem Bearbeiten der Untersuchungsmaterialien (Prätest) und unmittelbar danach (Posttest) erfasst. Auf diesem Wege sind die Autor:innen in der Lage, Veränderungsmessungen vorzunehmen und auszuschließen, dass die unmittelbar nach dem Lesen erfassten Unterschiede bereits vor der Rezeptionssituation bestanden.

Idealerweise würde das Design um einen Messzeitpunkt ergänzt, der zeitlich deutlich verzögert liegt und Aussagen über die zeitliche Stabilität der Befunde zulässt (Verzögerter Posttest). So zielen viele kommunikative Maßnahmen darauf ab, möglichst nachhaltige Veränderungen herbeizuführen, die sich nicht bloß auf rasch verblassende Eindrücke beschränken. Zudem kennt die Lernpsychologie auch sog. Sleeper-Effekte, bei denen die Vorteile einer Lernform erst nach einer Weile sichtbar werden (Köller 2015). Dies ist zum Beispiel der Fall, wenn nicht die Teilnahme an einer Maßnahme der Wissenschaftskommunikation selbst, sondern die in der Folge auftretende vermehrte Auseinandersetzung mit dem Thema zu einem substanziellen Lernzuwachs führt. Solche Effekte würden ohne Hinzufügen eines verzögerten Posttests übersehen und die Wirksamkeit einer Maßnahme systematisch unterschätzen.

4 Zur Wahl eines geeigneten Benchmarks in Evaluationsstudien oder „Was macht eigentlich die Kontrollgruppe?"

Soll die Wirkung einer Maßnahme der Wissenschaftskommunikation evaluiert werden, geht damit in der Regel die Entscheidung über eine geeignete Kontrollgruppe einher. Eine Taxonomie der Handlungsmöglichkeiten bietet die Klassifikation von Hager (2008), die ihren Ursprung in der Evaluation pädagogisch-psychologischer Interventionen hat und zwischen isolierter, vergleichender und kombinierter Evaluation unterscheidet.

Bei der *isolierten Evaluation* geht es darum, einen grundsätzlichen Wirknachweis einer Maßnahme zu erbringen. Die zu testende Maßnahme wird experimentell mit einer anderen Intervention von ähnlicher Komplexität und Dauer verglichen, die jedoch andere Ziele verfolgt. Beispielsweise könnten die Effekte eines Schülerlaborworkshops zu den Ursachen des Klimawandels mit einem in Dauer, Didaktik und Zielgruppe vergleichbaren Schülerlaborworkshop zu einem anderen Thema verglichen werden. Es würde erwartet, dass bei den Proband:innen der Interventionsgruppe, nicht jedoch bei denen der Kontrollgruppe, Veränderungen in der abhängigen Variable (z. B. Wissen über Ursachen des Klimawandels) auftreten. Hager (2008) spricht nur dann von einer „echten" Kontrollgruppe, wenn diese ebenfalls eine Maßnahme durchläuft. Ist dies nicht der Fall, kann nicht ausgeschlossen werden, dass Veränderungen in der abhängigen Variable ein methodisches Artefakt sind, das auf der erfahrenen pädagogischen Zuwendung und nicht auf den spezifischen Lernhandlungen beruht.

Bei der *vergleichenden Evaluation* werden zwei oder mehr konkurrierende Maßnahmen, die mit unterschiedlichen Mitteln dieselben Ziele erreichen wollen, hinsichtlich ihrer Wirksamkeit verglichen. Diese Variante testet folglich, ob eine neue Maßnahme ähnlich gute oder gar bessere Wirkungen erzielt als eine bereits bekannte, etablierte Maßnahme. Dieses Vorgehen bietet sich insbesondere dort an, wo mit einer isolierten Evaluation bereits ein Wirknachweis einer Maßnahme erbracht worden ist. So könnte zum Beispiel getestet werden, ob ein durch die Nutzung digitaler Apps unterstützter Museumsbesuch eine größere Interessenförderung bei jungen Museumbesucher:innen hervorruft als der alleinige Museumsbesuch. Vergleichenden Evaluationen können grundsätzlich drei Hypothesen zugrunde liegen: Bei der Äquivalenzhypothese erwarten die Forschenden, dass die zu testende Maßnahme genauso gute Ergebnisse erzielt, wie eine bereits etablierte (aber vielleicht aufwendigere) kommunikative Maßnahme. Die Überlegenheitshypothese geht hingegen davon aus, dass mit der zu testenden Maßnahme bessere Ergebnisse erzielt werden als mit der etablierten Variante, während die „Nicht-Unterlegenheitshypothese" annimmt, dass die zu testende Maßnahme mindestens ebenbürtig ist.

Die *kombinierte Evaluation* führt schließlich beide Evaluationsformen zusammen. Sie kommt zum Beispiel zum Einsatz, wenn sich in einer vergleichenden Evaluation zunächst zwei Interventionen als gleichwertig erwiesen haben und die Wirkungsfrage nun in einer Folgestudie unter Verwendung des Designs der isolierten Evaluation beantwortet werden soll.

5 Ausgewählte methodische Herausforderungen bei der Evaluation von Maßnahmen der Wissenschaftskommunikation mit Experimenten

Eine qualitativ hochwertige Evaluation von Maßnahmen der Wissenschaftskommunikation geht mit einer Vielzahl methodischer Herausforderungen einher, die eine sorgfältige Studienplanung unausweichlich machen. Hierzu gehört neben der Wahl eines geeigneten Studiendesigns auch der Umgang mit fehlenden Daten, die aufgrund des Drop-Outs von Versuchspersonen in Designs mit mehreren Messzeitpunkten ein besonderes Augenmerk verlangen (Köller 2015). Eine weitere Herausforderung ist die geschachtelte Datenstruktur in Studien, an denen Versuchspersonen in intakten Gruppen (z. B. Schulklassen in Schülerlaborworkshops) teilnehmen. Die Datenstruktur gilt es bei der Auswahl geeigneter statistischer Verfahren wie der Berechnung von Intraklassenkoeffizienten oder Mehrebenenanalysen zu berücksichtigen (Lüdtke und Köller 2010). Weil in

diesem Beitrag die experimentellen Methoden der Evaluation im Vordergrund stehen, liegt der Fokus der methodischen Überlegungen im Folgenden auf der Unterscheidung von experimentellen und quasiexperimentellen Ansätzen und dem mit quasiexperimentellen Ansätzen verbundenen Problem der Sicherung der internen Validität.

5.1 Was ist der Unterschied zwischen Experimenten und Quasiexperimenten?

Das Kernmerkmal echter Experimente ist die zufällige (fachsprachlich: "randomisierte") Zuteilung von Untersuchungsteilnehmer:innen zu Versuchsbedingungen, sodass jede Versuchsperson die gleiche Chance hat, in Experimental- oder Kontrollbedingung zu gelangen. Die Randomisierung sichert bei hinreichend großer Stichprobe, dass sich die Versuchsbedingungen nur in Bezug auf die manipulierte Variable, nicht jedoch in Bezug auf personenbezogene Drittvariablen wie Alter, Bildungshintergrund oder Themeninteresse unterscheiden.

Bei vielen Evaluationsstudien, die unter natürlichen Bedingungen (man sagt auch „im Feld") durchgeführt werden, dürfte die Randomisierung allerdings eine große Herausforderung darstellen[1]. Schließlich entscheiden Bürger:innen selbst, ob sie an einem Citizen-Science-Projekt teilnehmen, oder welchen Bürger:innendialog sie besuchen. Auch durch Ausstellungen in Science-Centern schlendern die Menschen üblicherweise in Gruppen (Lewalter und Noschka-Roos 2009) und können demzufolge nicht einfach durch die Experimentalleitung zufällig auf verschiedene Varianten der Ausstellung verteilt werden. Studien, in denen eine unabhängige Variable manipuliert wird, ohne dass eine randomisierte Zuteilung von Versuchspersonen zu Versuchsbedingungen erfolgt, werden als Quasiexperiment bezeichnet. Diese sind im Feld zwar leichter zu realisieren als reine Experimente, sie kommen aber oft mit der methodischen Einschränkung der Konfundierung von Versuchsbedingungen mit personenbezogenen Drittvariablen daher. Dies schränkt die interne Validität der Studie ein, weil Bedingungsunterschiede in den Ausprägungen der abhängigen Variable nicht mehr eindeutig auf die experimentelle Manipulation zurückgeführt werden können.

[1] Es sei jedoch darauf hingewiesen, dass die Forschung zur Wissenschaftskommunikation auch zahlreiche „reine" Experimente hervorgebracht hat, in denen Forscher:innen unter kontrollierten (Labor-)Bedingungen Gesetzmäßigkeiten nachspüren (z. B. Anderson et al. 2014; Bearth et al. 2022; Hendriks et al. 2016).

5.2 Maßnahmen zur Sicherung der internen Validität in Quasieexperimenten

In Quasiexperimenten gilt es bereits vor Studiendurchführung verschiedene Maßnahmen zu ergreifen, um die interne Validität zu erhöhen (für einen umfassenden Überblick, siehe Döring und Bortz 2016):

- *Erfassung konfundierender Variablen:* Um Konfundierungen überhaupt aufzudecken, ist es unausweichlich, dass Forschende vor Studienbeginn potenziell konfundierende Variablen identifizieren und diese in ihrer Studie ebenfalls erfassen. Die Identifikation solcher Variablen erfordert sowohl theoretische Überlegungen als auch Kenntnisse der einschlägigen empirischen Befundlage. Nur wenn konfundierende Variablen bekannt sind und erfasst wurden, können diese zum Beispiel mit der Kovarianzanalyse statistisch kontrolliert werden. Hierbei wird regressionsanalytisch geschätzt, welchen Wert die Versuchspersonen auf der abhängigen Variable einnähmen, wenn alle Proband:innen denselben Wert auf der Kovariate hätten (Döring und Bortz 2016). Für die anschließenden Mittelwertvergleiche werden sodann die geschätzten, um Unterschiede auf der Kovariaten bereinigten Werte verwendet.
- *Einführung eines Prätests:* Um auszuschließen, dass etwaige Bedingungsunterschiede nicht bereits vor Beginn der Studie existierten, können Forscher:innen zusätzlich zum Posttest einen Prätest durchführen. Im Idealfall unterscheiden sich die Bedingungen im Prätest nicht, wohingegen im Posttest Bedingungsunterschiede in der erwarteten Richtung auftreten. Die Ausgangswerte können bei der Datenanalyse auf verschiedene Weise berücksichtigt werden. Fließen sie als weiterer Messzeitpunkt in eine gemischte Varianzanalyse (siehe oben) ein, würde sich der erwartete Effekt in einer Interaktion zwischen dem Bedingungsfaktor und dem Messwiederholungsfaktor äußern. Eine weitere Möglichkeit besteht darin, den Effekt des Ausgangsniveaus statistisch mit der Methode der Kovarianzanalyse zu kontrollieren. Eine vergleichende Diskussion der verschiedenen analytischen Vorgehensweisen bietet Hager (2008).
- *Parallelisieren von Untersuchungsbedingungen:* Sind potenziell konfundierende Variablen bekannt, sollten diese nach Möglichkeit über alle Untersuchungsbedingungen konstant gehalten werden oder zumindest ähnlich ausgeprägt sein. Wer zum Beispiel eine Studie zur Wirkung authentischer Ausstellungsobjekte in einem Museum macht, sollte die Datenerhebungen stets zu denselben Wochentagen durchführen, um sicherzustellen, dass Faktoren

wie Besucher:innenstruktur oder Besucher:innendichte in allen Bedingungen vergleichbar ausgeprägt sind. In ähnlicher Weise könnten Forscher:innen, die Workshops in einem Schülerlabor durchführen, vorab vorliegende Informationen über die Schulform und den Sozialindex der entsendenden Schule nutzen, um sicherzustellen, dass Variablen wie Lernvoraussetzungen und sozioökonomische Herkunft über die Versuchsbedingungen hinweg vergleichbar ausgeprägt sind.

6 Fazit

Die Evaluation von Maßnahmen der Wissenschaftskommunikation ist ein wichtiges Instrument zur Qualitätssicherung solcher Initiativen und zugleich Ausdruck der zunehmenden Professionalisierung eines ‚boomenden' Tätigkeitsfeldes. Sie liefert wichtige Erkenntnisse über die Randbedingungen gelingender Wissenschaftskommunikation und erfüllt damit nicht nur eine innerwissenschaftlich relevante, sondern auch eine gesellschaftlich bedeutsame Funktion.

Im vorliegenden Beitrag lag der Fokus auf experimentellen Herangehensweisen in der Evaluation von Maßnahmen der Wissenschaftskommunikation. Diese stellen einerseits hohe methodische Anforderungen an die Studienplanung und -durchführung, bieten andererseits aber das Potenzial, belastbare Aussagen über Kausalzusammenhänge zu liefern. Dies mag überall dort erstrebenswert sein, wo Forschende vom Konkreten abstrahieren und Aussagen mit einem weiteren Geltungsbereich treffen wollen. Es geht dann nicht mehr primär um die Rekonstruktion der subjektiven Eindrücke von zufällig zu einem bestimmten Zeitpunkt an einem Ort befindlichen Personen. Diese sind in der hier vorgestellten Forschungstradition vielmehr Stellvertreter:innen einer größeren Grundgesamtheit, fachsprachlich eine Stichprobe aus einer Population, aus der man beliebig viele weitere Stichproben ziehen könnte. Das Ziel ist, Gesetzmäßigkeiten zu erforschen, die in einem genauer zu explizierenden Anwendungsbereich Gültigkeit besitzen und im Idealfall zu replizierbaren Effekten führen, was wiederum einen großen praktischen Nutzen hat.

Beim Design experimenteller Studien, die einen solchen Erkenntnisgewinn versprechen, können Evaluator:innen mittlerweile auf einen großen Fundus an Erkenntnissen und Best-Practice-Beispielen zurückgreifen, die in den letzten Jahrzehnten in benachbarten Disziplinen wie der empirischen Bildungsforschung gewonnen wurden. Hieraus sind verschiedene Lehrbücher und zahlreiche Fachartikel hervorgegangen, in denen das hier vorgestellte Kernproblem in der Regel deutlich ausführlicher behandelt wird, als es im Kontext dieses

Herausgeber:innenbands möglich und sinnvoll ist. Einige dieser Werke sind im vorliegenden Beitrag referenziert und bieten der interessierten Leser:innenschaft damit Anknüpfungspunkte, um tiefer in das Thema der Evaluation mit experimentellen Ansätzen einzusteigen.

Literatur

Anderson AA, Brossard D, Scheufele DA, Xenos MA, Ladwig P (2014) The "nasty effect:" online incivility and risk perceptions of emerging technologies. J Comput-Mediat Commun 19(3):373–387. https://doi.org/10.1111/jcc4.12009

Bauer TK, Gigerenzer G, Krämer W (2014) Warum dick nicht doof macht und Genmais nicht tötet. Über Risiken und Nebenwirkungen der Unstatistik. Campus, Frankfurt

Bearth A, Kaptan G, Kessler SH (2022) Genome-edited versus genetically-modified tomatoes: an experiment on people's perceptions and acceptance of food biotechnology in the UK and Switzerland. Agric Hum Values 39(3):1117–1131. https://doi.org/10.1007/s10460-022-10311-8

Bonney R, Phillips TB, Ballard HL, Enck JW (2015) Can citizen science enhance public understanding of science? Public Underst Sci 25(1):2–16. https://doi.org/10.1177/0963662515607406

Döring N, Bortz, J (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften (5. Aufl.). Springer, Berlin. https://doi.org/10.1007/978-3-642-41089-5

Field A, Hole G (2003) How to design and report experiments. Sage, London. https://doi.org/10.7748/nr.11.1.83.s13

Hager W (2008) Evaluation von pädagogisch-psychologischen Interventionsmaßnahmen. In: Schneider W, Hasselhorn M (Hrsg) Handbuch der Pädagogischen Psychologie. Hogrefe, Göttingen, S 721–732

Hendriks F, Kienhues D, Bromme R (2016) Disclose your flaws! Admission positively affects the perceived trustworthiness of an expert science blogger. Stud Commun Sci 16(2):124–131. https://doi.org/10.1016/j.scoms.2016.10.003

Köller O (2015) Evaluation pädagogisch-psychologischer Maßnahmen. In: Wild E, Möller J (Hrsg) Pädagogische Psychologie. Springer, S 329–342. https://doi.org/10.1007/978-3-642-41291-2_14

Lewalter D, Noschka-Roos A (2009) Museum und Erwachsenenbildung. In: Tippelt R, von Hippel A (Hrsg) Handbuch Erwachsenenbildung/Weiterbildung. VS Verlag, Wiesbaden, S 527–541. https://doi.org/10.1007/978-3-531-91834-1_32

Lin S-F, Lin H-S, Lee L, Yore LD (2015) Are science comics a good medium for science communication? The case for public learning of nanotechnology. Int J Sci Educ Part B 5(3):276–294. https://doi.org/10.1080/21548455.2014.941040

Lüdtke O, Köller O (2010) Mehrebenenanalyse. In: Rost DH (Hrsg) Handwörterbuch Pädagogische Psychologie (4. Aufl). Beltz/PVU, Weinheim, S 530–535

Scharrer L, Rupieper Y, Stadtler M, Bromme R (2017) When science becomes too easy: science popularization inclines laypeople to underrate their dependence on experts. Public Underst Sci 26(8):1003–1018. https://doi.org/10.1177/0963662516680311

Schwan S, Bauer D, Kampschulte L, Hampp C (2016) Representation equals presentation? Photographs of objects receive less attention and are less well remembered than real objects. J Media Psychol 29(4):176–187. https://doi.org/10.1027/1864-1105/a000166

Marc Stadtler ist Professor für Kompetenzentwicklung und Kompetenzmodellierung am Institut für Erziehungswissenschaft der Ruhr-Universität Bochum. In seiner Forschung beschäftigt er sich mit Fragen der kritischen Nutzung von Wissenschaftsinformationen in digitalen Informationsumwelten, insbesondere mit der Förderung der Kompetenz zur kritischen Bewertung von Wissenschaftsinformationen bei Kindern, Jugendlichen und jungen Erwachsenen.

Corinna Schuster ist Akademische Rätin a. Z. im Arbeitsbereich Kompetenzentwicklung und Kompetenzmodellierung am Institut für Erziehungswissenschaft der Ruhr-Universität Bochum. In ihrer Forschung beschäftigt sie sich mit dem Transfer metakognitiver Strategien beim selbstregulierten Lernen sowie der Akkuratheit von metakognitiven Verstehensurteilen beim Lesen in digitalen Informationsumwelten.



Praxisbeitrag: Experimentelle Methoden in der evaluativen Wissenschaftskommunikationsforschung am Beispiel von Bürgerwissenschaftsprojekten

Hannah Greving, Till Bruckermann und Joachim Kimmerle

Zusammenfassung

Effektive Wissenschaftskommunikation wird in der Gesellschaft immer wichtiger. Daher ist es umso entscheidender, zu verstehen, wann und unter welchen Umständen Wissenschaftskommunikation gelingen kann. Eine anerkannte Methode dafür ist die experimentelle Vorgehensweise. Experimente erlauben es, kontrollierte Untersuchungen durchzuführen und Ursache-Wirkungsbeziehungen zu überprüfen. Sie können daher einen entscheidenden Beitrag zur Evaluation von Wissenschaftskommunikation leisten. In diesem Beitrag verdeutlichen wir anhand von drei Beispielen aus der Praxis, dass Experimente erfolgreich für die Erforschung von Wissenschaftskommunikation und genauer von Bürgerwissenschaftsprojekten ein- und umgesetzt werden können. Vor allem dann, wenn Forschende und Durch-

H. Greving (✉) · J. Kimmerle
Arbeitsgruppe Wissenskonstruktion, Leibniz-Institut für Wissensmedien, Tübingen, Deutschland
E-Mail: h.greving@iwm-tuebingen.de

J. Kimmerle
E-Mail: j.kimmerle@iwm-tuebingen.de

T. Bruckermann
Institut für Erziehungswissenschaft, Leibniz Universität Hannover, Hannover, Deutschland
E-Mail: till.bruckermann@iew.uni-hannover.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation*, https://doi.org/10.1007/978-3-658-39582-7_19

führende von Bürgerwissenschaftsprojekten beabsichtigen, die Wirksamkeit ihres Projektes zu evaluieren, stellen Experimente eine effektive Methode dar.

Effektive Wissenschaftskommunikation bekommt einen immer größeren Stellenwert in der Gesellschaft. Umso entscheidender ist es daher zu verstehen, wann und unter welchen Umständen Wissenschaftskommunikation gelingen kann und welche Faktoren zu ihrem Erfolg beitragen können. Eine Methode, die in der Psychologie und anderen Disziplinen bereits eine lange Tradition hat, ist die experimentelle Vorgehensweise. Experimente erlauben es, kontrollierte Untersuchungen durchzuführen und Ursache-Wirkungsbeziehungen zu überprüfen, indem ein angenommener ursächlicher Faktor in unterschiedlichen Gruppen variiert wird, um dessen Wirkung auf eine oder mehrere Zielvariablen zu untersuchen. Alle anderen möglichen Einflussfaktoren werden zugleich vollständig konstant gehalten. Wenn es Unterschiede hinsichtlich der Zielvariable zwischen den Gruppen gibt, ist dieser Unterschied daher allein auf die Variation des angenommenen ursächlichen Faktors zurückzuführen (vgl. Döring und Bortz 2016). Diese Logik experimenteller Forschung kann zur Evaluation von Wissenschaftskommunikation einen entscheidenden Beitrag leisten. In diesem Beitrag beschäftigen wir uns daher mit Beispielen aus der Praxis, in denen experimentelle Methoden in der Forschung zur Wissenschaftskommunikation erfolgreich eingesetzt wurden. Im Folgenden gehen wir zunächst kurz auf Experimente und deren Vor- und Nachteile ein. Anschließend wenden wir uns Bürgerwissenschaften zu, die teilweise als partizipative Wissenschaftskommunikation verstanden werden können, und gehen auf erfolgreiche Praxisbeispiele (d. h. in unserem Fall erfolgreich durchgeführte experimentelle Studien) in diesem Bereich ein.

1 Experimente und ihre Vor- und Nachteile

Experimente haben in der Medizin, der Psychologie, der Ökonomie und weiteren empirisch orientierten Sozialwissenschaften eine lange Tradition und folgen dabei einer eingehenden Logik, die wir nun anhand eines einfachen Beispiels verdeutlichen: Evaluierende im Bereich der Wissenschaftskommunikationsforschung (d. h. Forschende sowie gleichermaßen Praktiker:innen) möchten gerne wissen, ob ein wissenschaftsjournalistischer Text, der auf blauem Hintergrund geschrieben wurde, als vertrauenswürdiger wahrgenommen wird als ein ebensolcher Text vor rotem Hintergrund (z. B. Mehta und Zhu 2009). Um diese Frage zu untersuchen, wird ein und derselbe neutral geschriebene Text einer

Gruppe von Personen auf blauem und einer anderen Gruppe von Personen auf rotem Hintergrund präsentiert. Wichtig ist dabei, dass die Personen zufällig der einen oder anderen Gruppe zugewiesen wurden, sodass sich die Personen in den beiden Gruppen nicht systematisch voneinander unterscheiden. Danach werden beide Gruppen gefragt, als wie vertrauenswürdig sie den Text empfunden haben. Das heißt, der Faktor, der variiert wird, ist die Hintergrundfarbe (blau vs. rot). Alle anderen Aspekte werden über beide Gruppen hinweg konstant gehalten, etwa der Inhalt oder die Länge des Textes. Die Zielvariable, die gemessen wird, ist die empfundene Glaubwürdigkeit. Wenn sich die blaue Gruppe bei der Auswertung der Daten nun tatsächlich in statistisch bedeutsamer Weise in ihrer Einschätzung der Glaubwürdigkeit von der roten Gruppe unterscheidet, dann spricht das für die vorherige Vermutung.

Dieses Beispiel beschreibt die einfachste Form eines Experimentes und ist auf verschiedene Art und Weise erweiterbar. So könnten beispielsweise die Stufen des Faktors Hintergrundfarbe erweitert und zusätzlich die Wirkung von grün und gelb als Hintergrundfarbe getestet werden. Es könnte ein weiterer Faktor hinzugefügt und überprüft werden, ob sich beispielsweise die zusätzliche Präsentation eines zum Text passenden Fotos positiv auf die Glaubwürdigkeit auswirkt. Außerdem ließen sich neben der Zielvariable Glaubwürdigkeit weitere Variablen untersuchen, wie beispielsweise empfundene positive Emotionen oder die Erinnerung an Informationen aus dem wissenschaftsjournalistischen Text. Schlussendlich könnte der Faktor Zeit berücksichtigt werden und der Effekt der Hintergrundfarbe über einen längeren Zeitraum hinweg getestet werden.

Experimente haben verschiedene Vor- und Nachteile, die Evaluierende berücksichtigen sollten, wenn sie sich für oder gegen ein experimentelles Vorgehen entscheiden. Zu den Vorteilen zählen die kontrollierten Bedingungen, unter denen Experimente stattfinden. Sie erlauben es den Evaluierenden, klare Aussagen über Ursache-Wirkungsbeziehungen zu treffen, da der Einfluss von potenziellen Störvariablen oder verzerrenden Variablen relativ gering ist. Es gibt allerdings bestimmte Voraussetzungen, damit Experimente in kontrollierter Weise ablaufen können wie beispielsweise die zufällige Zuweisung der Teilnehmenden zu den verschiedenen Bedingungen.

Zu den Nachteilen zählt, dass Ergebnisse von Experimenten durch die kontrollierten Bedingungen, unter denen sie stattfinden, oftmals nur eine geringe Aussagekraft hinsichtlich der Wirkung in echten Situationen haben. Unser Alltagsleben ist in der Regel von einer hohen Komplexität geprägt und es treten viele verschiedene Variablen (darunter auch potenzielle Störvariablen und verzerrende Variablen) gleichzeitig auf. Ein Experiment ist in der Regel nicht in der Lage diese Komplexität vollständig abzubilden, sodass die Generalisierbarkeit

von in Experimenten gefundenen Ergebnissen eingeschränkt ist. Ein Experiment ist ebenfalls auch nur in der Lage eine sehr begrenzte Anzahl von Wirkfaktoren (z. B. Hintergrundfarbe, Präsentation zusätzlicher Fotos) auf eine begrenzte Anzahl von Zielvariablen (z. B. Vertrauenswürdigkeit, positive Emotionen, Erinnerung) zu untersuchen. Alle anderen, darüber hinausgehenden Faktoren oder Variablen können nicht berücksichtigt werden, sodass unter Umständen nur ein unvollständiges Bild entsteht (für weitere Details zu Experimenten siehe auch Stadtler und Schuster in diesem Band).

Diese Vor- und Nachteile sollten Evaluierende im Bereich der Wissenschaftskommunikationsforschung je nach Forschungsfrage und Möglichkeiten der methodischen Umsetzung bei der Entscheidung für ein Experiment gegeneinander abwägen. Aus unserer Sicht stellen die kontrollierten Bedingungen von Experimenten und ihre Fähigkeit, Ursachen und Wirkungen klar zu benennen, einen erheblichen Vorteil dar, gerade in Bereichen, in denen bisher wenig zu Ursache-Wirkungsbeziehungen geforscht wurde.

2 Bürgerwissenschaften als partizipative Wissenschaftskommunikation

Wissenschaftskommunikation stellt ein breites Feld dar und rückt immer stärker in den Fokus des medialen und öffentlichen Interesses. In den letzten Jahrzehnten hat sich Wissenschaftskommunikation von einer unidirektionalen Vermittlung wissenschaftlicher Erkenntnisse, in der Wissenschaftler:innen oder Wissenschaftsjournalist:innen an die Öffentlichkeit kommunizieren, zu einem multidirektionalen Austausch gewandelt, in dem sich Wissenschaftler:innen und die Öffentlichkeit in partizipativen Formaten gleichermaßen an Wissenschaft beteiligen (z. B. Ball 2020). Über die Wirkung solcher Formate ist bisher jedoch nur wenig bekannt. Insbesondere geht es in diesem Beitrag um die Wirkung von Projekten aus dem Bereich der sogenannten Bürgerwissenschaften, mit deren erfolgreicher Gestaltung im Sinne gelungener Wissenschaftskommunikation wir uns im Folgenden beschäftigen.

Unter Bürgerwissenschaften wird die Zusammenarbeit zwischen engagierten Bürger:innen und professionell arbeitenden Wissenschaftler:innen in vorwiegend wissenschaftlichen Projekten verstanden (Heigl et al. 2019; für einen Überblick siehe Vohland et al. 2021). Obwohl Bürger:innen in vielen Projekten vor allem die Datensammlung von Wissenschaftler:innen unterstützen, gibt es weitere Ansätze, bei denen die Zusammenarbeit zwischen Bürger:innen und Wissenschaftler:innen als kollaborativer Prozess (z. B. Shirk et al. 2012;

siehe auch Cress und Kimmerle 2018) und Wissensaustausch verstanden wird (z. B. Bruckermann et al. 2021; Greving et al. 2022). Diesen Ansätzen folgend beteiligen sich beide Gruppen von Beteiligten an einem Dialog über die wissenschaftlichen Erkenntnisse und haben so die Möglichkeit, voneinander zu lernen. Demzufolge können insbesondere kollaborative Projekte in den Bürgerwissenschaften als eine partizipative Form der Wissenschaftskommunikation verstanden werden.

Bürgerwissenschaftsprojekte gewinnen zunehmend an Bedeutung, sowohl auf nationaler als auch auf internationaler Ebene (z. B. Vohland et al. 2021). Vorreiter in der deutschen Community der Bürgerwissenschaften sind zum Beispiel die Initiativen und Plattformen von Bürger schaffen Wissen (https://www.buergerschaffenwissen.de/) und Wissenschaft im Dialog (https://www.wissenschaft-im-dialog.de/), die die Vernetzung von Bürgerwissenschaftsprojekten und den Austausch zwischen Wissenschaft und Öffentlichkeit unterstützen. In der breiten Auswahl von internationalen Bürgerwissenschaftsplattformen gibt es beispielsweise die Plattform Zooniverse (https://www.zooniverse.org/), auf der über 300 Projekte aus den unterschiedlichsten Bereichen wie Astronomie, Kunst, Geschichte oder Medizin angeboten werden und an denen über 2,4 Mio. Nutzer:innen beteiligt sind, die bereits mehr als 679 Mio. Objekte klassifiziert haben.

3 Praxisbeispiele experimenteller Untersuchungen im Bereich der Bürgerwissenschaften

Trotz des Erfolgs von Bürgerwissenschaften wurden in der Praxis bisher wenige systematische Evaluationen von Bürgerwissenschaftsprojekten durchgeführt. Wenn Evaluationen durchgeführt oder Teile eines Projektes näher untersucht wurden, dann wurden vor allem auf deskriptiver Ebene Fallbeispiele einzelner Projekte beschrieben, ohne dass diese Projekte näher systematisch und empirisch untersucht wurden. Außerdem wurden Effekte auf die Teilnehmenden nur selten untersucht (z. B. Aristeidou und Herodotou 2020) und, wenn doch, wurden selten experimentelle Methoden dafür ausgewählt. Im Folgenden gehen wir daher auf drei Beispiele aus der Praxis ein, bei denen experimentelle Untersuchungen im Bereich der Bürgerwissenschaften durchgeführt wurden. In diesen Beispielen unterscheiden wir zwischen psychologischen Studien im Labor und im Feld. Laborstudien finden in der Regel in ruhigen, reizarmen Umgebungen und eigens dafür vorgesehenen Laborräumen statt, in denen Versuchsteilnehmende zumeist allein an einem Computer Aufgaben erledigen. Feldstudien hingegen finden in

der Regel in einem natürlichen Kontext und einer alltäglichen Umgebung der Teilnehmenden statt und werden bevorzugt durch Fragebögen erhoben.

3.1 Praxisbeispiel 1: Experimente mit Bürger:innen eines Bürgerwissenschaftsprojekts im Feld

Im ersten Beispiel wurden experimentelle Untersuchungen mit Bürger:innen eines realen Bürgerwissenschaftsprojekts im Feld durchgeführt (Greving et al. 2022). Im Rahmen von vier Feldstudien wurden Bürger:innen begleitet, die an einem Bürgerwissenschaftsprojekt zum Thema Fledermäuse in der Stadt Berlin teilnahmen. Die generelle Fragestellung in diesen Studien war, ob sich durch die Beteiligung am Bürgerwissenschaftsprojekt über die Zeit hinweg Veränderungen in den Einstellungen zu Fledermäusen, dem Fachwissen über Fledermäuse und den Einstellungen, sich für Bürgerwissenschaftsprojekte zu engagieren, ergeben. Mit einem experimentellen Ansatz wurde genauer untersucht, ob im Vergleich zur reinen Beteiligung an der Datensammlung die zusätzliche Beteiligung an der Datenanalyse einen positiveren Effekt auf die Einstellungen und das Fachwissen hat.

Um diese Fragestellungen zu beantworten, wurde ein sogenanntes Prä-Post-Design gewählt, über das der Grad der Beteiligung als Faktor variiert wurde. Dazu füllten zunächst alle teilnehmenden Bürger:innen zu Beginn des Bürgerwissenschaftsprojekts einen Prä-Fragebogen aus, der Einstellungen zu Fledermäusen, Fachwissen über Fledermäuse und Einstellungen zur Beteiligung an Bürgerwissenschaften erfasste. Alle Teilnehmenden beteiligten sich danach zunächst an der Datensammlung des Projekts. Die Datensammlung bestand darin zweimal eine von den beteiligten Wissenschaftler:innen vordefinierte Route abzulaufen und währenddessen die Rufe von in der Luft fliegenden Fledermäusen mit einem Detektor zu erfassen. Diese Fledermausrufe luden die Teilnehmenden dann auf einer Plattform hoch, wo die beteiligten Wissenschaftler:innen die zu den Rufen gehörenden Fledermausarten bestimmten. Nach der Datensammlung füllte in etwa die Hälfte der Teilnehmenden, die zufällig ausgewählt wurden, den Post-Fragebogen aus, der mit dem Prä-Fragebogen identisch war. Danach beteiligten sich alle Teilnehmenden an einer strukturierten Datenanalyse, die ebenfalls auf der Plattform stattfand. Sie bestand darin, das Vorkommen bestimmter Fledermausarten mit verschiedenen Umweltvariablen (z. B. künstliches Licht, dichter Baumbestand, viele Wasserstellen) in Beziehung zu setzen. Diesen Analyseprozess konnten die Teilnehmenden sowohl für ihre eigenen gesammelten Daten als auch für die Daten aller Teilnehmenden durchführen. Nach Abschluss dieser

Analyse füllte schlussendlich die andere Hälfte der Teilnehmenden den Post-Fragebogen aus.

Die Ergebnisse der Studie zeigten, dass die Teilnehmenden während ihrer Beteiligung an dem Projekt eine positivere Einstellung zu Fledermäusen, ein höheres Fachwissen zu Fledermäusen und positivere Einstellungen zur Beteiligung an Bürgerwissenschaftsprojekten entwickelten. Es zeigte sich allerdings auch, dass es für diese positiven Entwicklungen unerheblich war, ob die Teilnehmenden nur Daten gesammelt hatten oder Daten gesammelt und zusätzlich auch analysiert hatten. Weitere Untersuchungen müssten die Voraussetzungen und Bedingungen bestimmen, unter denen die Datenanalyse positive Effekte auf die Teilnehmenden hat. Ebenso könnten solche Untersuchungen auch andere Stichproben testen, da an diesem Praxisbeispiel überwiegend gut gebildete und sehr an Fledermäusen interessierte Bürger:innen teilgenommen hatten. Zusammengefasst konnte dieses erste Praxisbeispiel dazu beitragen, zu verstehen, ob die zusätzliche Beteiligung von Bürger:innen an der Datenanalyse einen Mehrwert gegenüber der reinen Beteiligung an der Datensammlung mit sich bringt. Diese Erkenntnisse sind für Evaluierende (d. h. Forschende wie Praktiker:innen) im Bereich der Bürgerwissenschaften relevant, da das reine Angebot Daten zu analysieren vermutlich nicht ausreicht, sondern Bürger:innen dafür mehr Anleitung benötigen, damit die Datenanalyse einen positiven Effekt auf sie ausübt.

3.2 Praxisbeispiel 2: Experimente mit Studierenden im Labor zur Beteiligung an Bürgerwissenschaftsprojekten

Im zweiten Beispiel haben wir experimentelle Laborstudien mit Studierenden durchgeführt (Greving et al. 2020). Hierbei wurde die Frage untersucht, welche typischen Beteiligungsschritte in einem Bürgerwissenschaftsprojekt zu Wildtieren sich positiv auf das Gefühl des psychologischen Ownership auswirken. Mit psychologischem Ownership ist das subjektive Gefühl gemeint, etwas persönlich zu besitzen (Pierce et al. 2003). Dabei kann es sich um etwas Konkretes wie einen Stift oder eine Tasse handeln (Peck und Shu 2009). Es kann aber auch etwas Abstraktes wie eine Idee, das Unternehmen, für das man arbeitet (Van Dyne und Pierce 2004), oder eben ein Bürgerwissenschaftsprojekt sein. In der Untersuchung ging es um ein Projekt zu Wildtieren. In solchen Projekten werden Bürger:innen typischerweise mit Wildtierkameras ausgestattet, die in der Lage sind, unbemerkt Fotos von vorbeilaufenden Wildtieren aufzunehmen.

Außerdem ist es typischerweise die Aufgabe der Bürger:innen, die Fotos auf der Speicherkarte der Wildtierkamera in eine Datenbank (zu der auch die beteiligten Wissenschaftler:innen Zugriff haben) hochzuladen und danach die Wildtiere auf den Fotos zu bestimmen.

Entlang dieser typischen Aufgaben wurden die von den Teilnehmenden zu erledigenden Aufgaben zwischen verschiedenen Gruppen experimentell untersucht. Variiert wurde dabei, ob die Teilnehmenden eine Wildtierkamera in der Hand hatten oder nicht, ob sie der Kamera eine Speicherkarte entnommen und die Fotos auf der Speicherkarte auf einen Computer hochgeladen haben oder nicht und ob die Teilnehmenden die Wildtiere auf den Fotos der Speicherkarte bestimmen mussten oder nicht. Zusätzlich wurden die Teilnehmenden instruiert, die Aufgaben so zu erledigen als würden sie selbst an dem Projekt teilnehmen, als würden sie für eine andere Person teilnehmen oder sie erhielten keine diesbezügliche Instruktion. In zwei Studien wurde die Hypothese geprüft, dass die Teilnehmenden umso mehr Ownership empfinden sollten, je mehr Aufgaben sie erledigten.

Die Ergebnisse der Studien konnten diese Erwartung bestätigen. Das heißt, wenn Teilnehmende eine Wildtierkamera in der Hand hatten, mit der Speicherkarte interagiert haben, alle Fotos auf der Speicherkarte bestimmten und diese Aufgaben aus der eigene Perspektive heraus erledigten, dann empfanden sie mehr psychologisches Ownership als solche Teilnehmenden, die all diese Aufgaben nicht durchgeführt haben. Dabei gab es jedoch eine Ausnahme. Wenn Teilnehmende für eine andere Person teilnahmen, dabei aber alle Wildtiere auf den Fotos selbst bestimmen mussten, dann war das eher hinderlich für das empfundene Ownership. Im Nachhinein erklärten wir uns dies damit, dass solche Teilnehmenden womöglich das Gefühl hatten, dass das Bestimmen der Fotos nicht Teil der ihnen zugeschriebenen Rolle war. Aber diese Vermutung müsste in weiteren Studien noch geklärt werden. Darüber hinaus zeigten die Ergebnisse, dass die Einstellungen der Teilnehmenden zu Bürgerwissenschaften umso positiver und ihre Intention, sich auch zukünftig an Bürgerwissenschaftsprojekten zu beteiligen, umso stärker war, je mehr psychologisches Ownership sie empfanden. Diese Erkenntnisse belegen damit die wichtige Rolle von Ownership für Bürgerwissenschaften, müssten aber gleichzeitig zukünftig im Feld näher untersucht werden, da die Bedingungen im Labor nicht alle Merkmale von Bürgerwissenschaften im Feld abbilden können. Auch müsste geklärt werden, ob sich die positivere Einstellung zu und die stärkere Intention, sich an Bürgerwissenschaften zu beteiligen, auch tatsächlich im Verhalten zeigen lässt. Zusammengefasst zeigt dieses zweite Praxisbeispiel, dass experimentelle Laborstudien dabei helfen können herauszufinden, welche genauen Beteiligungsschritte

in einem Bürgerwissenschaftsprojekt zu Wildtieren förderlich (oder hinderlich) für das empfundene Ownership sind. Diese Ergebnisse zeigen auf, dass es sich für Praktiker:innen im Bereich der Wissenschaftskommunikationsforschung lohnt, in zukünftigen Bürgerwissenschaftsprojekten die Rolle von empfundenem Ownership mit zu berücksichtigen, weil es andere Variablen, wie beispielsweise die Einstellungen von Teilnehmenden, positiv beeinflussen kann.

3.3 Praxisbeispiel 3: Experimente mit Studierenden im Labor zu emotionalen Einflussfaktoren

Im dritten Beispiel haben wir ebenfalls experimentelle Laborstudien mit Studierenden durchgeführt (Greving und Kimmerle 2021). Während es im vorangegangenen Beispiel nur eine untergeordnete Fragestellung war, wie sich Ownership auf Einstellungen und Verhaltensintentionen bezüglich Bürgerwissenschaften auswirkt, wollten wir in diesem Beispiel konkret wissen, welche Faktoren sich positiv auf die Bereitschaft auswirken, sich in Bürgerwissenschaftsprojekten zu engagieren. Insbesondere standen Emotionen und ganz konkret die Emotion Mitgefühl im Fokus. Mitgefühl heißt, mit anderen mitzufühlen und sich in deren emotionalen Zustand hineinzuversetzen (Goetz et al. 2010). Somit ist Mitgefühl eine auf andere fokussierte Emotion und wird durch Sorge für andere geprägt. Sie geht mit Hilfeverhalten einher, wie beispielsweise erhöhter Spendenbereitschaft oder Engagement für den Klimaschutz, und ist daher auch für den Bereich der Bürgerwissenschaften von Interesse.

In zwei Studien haben wir daher die Rolle von Mitgefühl für Einstellungen zu Bürgerwissenschaften und Intentionen, sich an Bürgerwissenschaften zu beteiligen, untersucht. Auch in diesem Fall ging es um Bürgerwissenschaften zu Wildtieren, wobei sich eine Studie konkret mit Waschbären und die zweite Studie mit Füchsen beschäftigte. Dabei zeigten wir einer Gruppe von Studierenden Fotos von neutral dargestellten Waschbären oder Füchsen, einer zweiten Gruppe Fotos von bedrohlichen Waschbären oder Füchsen und einer dritten Gruppe Fotos von verletzten oder toten Waschbären oder Füchsen. In der Fuchsstudie zeigten wir einer vierten Gruppe von Studierenden außerdem Fotos von niedlichen Füchsen. Alle Studierenden wurden komplett zufällig einer dieser Gruppen zugewiesen. Danach erfassten wir die durch die Fotos ausgelösten Emotionen sowie die Einstellung zu Bürgerwissenschaften und die Intention, sich auch zukünftig an Bürgerwissenschaften zu beteiligen.

Die Ergebnisse der Studien zeigten, dass die Fotos von verletzten oder toten Waschbären oder Füchsen mehr Mitgefühl ausgelöst haben als die anderen

Fotos. Je mehr Mitgefühl die Teilnehmenden erlebten, desto positiver war ihre Einstellung zu Bürgerwissenschaften und desto eher waren sie auch bereit sich zukünftig an Bürgerwissenschaften zu beteiligen. Näher untersucht werden könnte auf Basis dieser Befunde, ob es entscheidend ist, warum die Wildtiere verletzt oder tot waren (d. h. ob es menschenverursacht war oder ob es natürliche Gründe hatte). Außerdem wäre es spannend zu untersuchen, wann sich der Effekt auf Mitgefühl umkehren würde (d. h. wie viele verletzte oder tote Wildtiere erträglich wären) und wann Bürger:innen eher gleichgültig reagieren würden. Zusammengefasst zeigen die Befunde des dritten Praxisbeispiels, dass Mitgefühl – in dosierten Maßen ausgelöst – eine entscheidende Rolle dabei spielen kann, Menschen für Bürgerwissenschaften sowie auch für den Wildtierschutz (Straka et al. 2021) zu begeistern. Für Durchführende von Bürgerwissenschaftsprojekten zum Artenschutz heißt das, dass der wohldosierte Einsatz von Bildmaterial von verletzten Wildtieren bei der Werbung für Projekte helfen kann.

4 Fazit und Ausblick

Experimentelle Methoden stellen eine sinnvolle Vorgehensweise dar, um klar und deutlich die Wirkung eines ursächlichen Faktors auf eine Zielvariable zu identifizieren. Sie können sowohl in Feldstudien (d. h. Studien in natürlichen, alltäglichen Umgebungen) als auch in Laborstudien (d. h. Studien in kontrollierten, reizarmen Umgebungen) zum Einsatz kommen und sind äußerst relevant für partizipative Formate der Wissenschaftskommunikation wie Bürgerwissenschaften. Bisher wurden Experimente nur selten zur Evaluation von Bürgerwissenschaftsprojekten hinsichtlich ihrer Effekte auf Bürger:innen eingesetzt. In diesem Beitrag haben wir anhand von Praxisbeispielen aufgezeigt, dass Experimente erfolgreich für die Erforschung von Wissenschaftskommunikation ein- und umgesetzt werden können. Somit konnten wir zeigen, dass Forschende und Durchführende von Bürgerwissenschaftsprojekten viel aus Experimenten lernen können, vor allem dann, wenn sie beabsichtigen die Wirksamkeit eines Projektes zu evaluieren. Beispielsweise scheint es nicht zu reichen, die Möglichkeit der Datenanalyse nur anzubieten. Praktiker:innen sollten Bürger:innen daher stärker bei der Datenanalyse anleiten, damit diese einen positiven Effekt auf die Bürger:innen haben kann. Ebenso sollten Durchführende von Projekten das empfundene Ownership der Bürger:innen sowie deren empfundene Rolle im Projekt mitberücksichtigen, da beides entscheidend zum Ausgang des Projektes für die Bürger:innen beitragen kann. Zuletzt kann auch nicht so schönes Bild-

material von verletzten oder toten Wildtieren sparsam eingesetzt zur Beteiligung an Bürgerwissenschaftsprojekten motivieren.

Die dargestellten Praxisbeispiele zeigen auf, dass die Bereiche der Bürgerwissenschaften und Wissenschaftskommunikation davon profitieren könnten, wenn Experimente im Vorfeld direkt mit eingeplant werden. Außerdem könnte es von Vorteil sein, etablierte Theorien und Konzepte bei der Planung und Umsetzung von Projekten zu berücksichtigen. Bei der Entscheidung für ein experimentelles Vorgehen sollten bekannte und bereits etablierte Forschungsdesigns verwendet und bei der Erhebung der Zielvariablen auf valide und reliable Maße geachtet werden. Schlussendlich könnte die Verwendung von experimentellen Methoden ein Qualitätskriterium für zukünftige Forschung im Bereich der Bürgerwissenschaften und der Wissenschaftskommunikation sein. Denn aus unserer Sicht kann der Einsatz von Experimenten absolut gewinnbringend sein.

Literatur

Aristeidou M, Herodotou C (2020) Online citizen science: a systematic review of effects on learning and scientific literacy. Citiz Sci 5(1):1–12. https://doi.org/10.5334/cstp.224

Ball R (2020) Wissenschaftskommunikation im Wandel: Von Gutenberg bis Open Science. Springer, Berlin

Bruckermann T, Greving H, Schumann A, Stillfried M, Börner K, Kimmig S, Hagen R, Brandt M, Harms U (2021) To know about science is to love it? Unraveling the knowledge-attitude relationship in Citizen Science on urban wildlife. J Res Sci Teach 58(8):1179–1202. https://doi.org/10.1002/tea.21697

Cress U, Kimmerle J (2018) Collective knowledge construction. In: Fischer F, Hmelo-Silver CE, Goldman SR, Reimann P (Hrsg) International handbook of the learning sciences. Routledge, New York, S 137–146

Döring N, Bortz J (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, 5. Aufl. Springer, Berlin

Goetz JL, Keltner D, Simon-Thomas E (2010) Compassion: an evolutionary analysis and empirical review. Psychol Bull 136(3):351–374. https://doi.org/10.1037/a0018807

Greving H, Kimmerle J (2021) You poor little thing! The role of compassion for wildlife conservation. Hum Dimens Wildl 26(2):115–131. https://doi.org/10.1080/10871209.2020.1800146

Greving H, Bruckermann T, Kimmerle J (2020) This is my project! The influence of involvement on psychological ownership and wildlife conservation. Curr Res Soc Psychol 1:100001. https://doi.org/10.1016/j.cresp.2020.100001

Greving H*, Bruckermann T*, Schumann A, Straka TM, Lewanzik D, Voigt-Heucke SL, Marggraf L, Lorenz J, Brandt M, Voigt CC, Harms U, Kimmerle J (2022) Improving attitudes and knowledge in a citizen science project on urban bat ecology. Ecol Soc 27(2):Article 24. (*shared first-authorship). https://doi.org/10.5751/es-13272-270224

Heigl F, Kieslinger B, Paul KT, Uhlik J, Dörler D (2019) Opinion: toward an international definition of citizen science. P Natl Acad Sci USA 116(17):8089–8092. https://doi.org/10.1073/pnas.1903393116
Mehta R, Zhu RJ (2009) Blue or red? Exploring the effect of color on cognitive task performances. Science 323(5918):1226–1229. https://doi.org/10.1126/science.1169144
Peck J, Shu SB (2009) The effect of mere touch on perceived ownership. J Consum Res 36(3):434–447. https://doi.org/10.1086/598614
Pierce JL, Kostova T, Dirks KT (2003) The state of psychological ownership: integrating and extending a century of research. Rev Gen Psychol 7(1):84–107. https://doi.org/10.1037/1089-2680.7.1.84
Shirk JL, Ballard HL, Wilderman CC, Phillips T, Wiggins A, Jordan R, McCallie E, Minarchek M, Lewenstein BV, Krasny ME, Bonney R (2012) Public participation in scientific research: a framework for deliberate design. Ecol Soc 17(2):29. https://doi.org/10.5751/ES-04705-170229
Straka TM, Greving H, Voigt CC (2021) The effects of bat photographs on emotions, attitudes, intentions, and wildlife value orientations. Hum Dimens Wildl 26(6):596–603. https://doi.org/10.1080/10871209.2020.1864068
Van Dyne L, Pierce JL (2004) Psychological ownership and feelings of possession: Three field studies predicting employee attitudes and organizational citizenship behavior. J Organ Behav 25(4):439–459. https://doi.org/10.1002/job.249
Vohland K, Land-Zandstra A, Ceccaroni L, Lemmens R, Perelló J, Ponti M, Samson R, Wagenknecht K (2021) The science of citizen science. Springer Nature, Berlin

Hannah Greving ist wissenschaftliche Mitarbeiterin am Leibniz-Institut für Wissensmedien in Tübingen. Sie beschäftigt sich mit dem Einfluss von Bürgerwissenschaftsprojekten auf die Teilnehmenden und effektiver Wissenschaftskommunikation.

Till Bruckermann ist Universitätsprofessor an der Leibniz Universität Hannover. Er forscht zu informellem Lernen in Bürgerwissenschaftsprojekten und insbesondere zur Entwicklung eines Wissenschaftsverständnisses.

Joachim Kimmerle ist stellvertretender Arbeitsgruppenleiter der Arbeitsgruppe Wissenskonstruktion am Leibniz-Institut für Wissensmedien in Tübingen sowie außerplanmäßiger Professor an der Eberhard Karls Universität Tübingen. Er interessiert sich für kollaborative Wissenskonstruktion, Gesundheitsverständnis und Wissenschaftskommunikation.



Praxisbeitrag: Kreative Feedbackmethoden zur Unterstützung von Prozessen und Veranstaltungen der Bürger:innen- und Öffentlichkeitsbeteiligung

Eva Wollmann und Jacob Birkenhäger

Zusammenfassung

Verfahren der Bürger- und Öffentlichkeitsbeteiligung beziehen oftmals wissenschaftliches Wissen und konkrete Forschungserkenntnisse ein, sind im Rahmen der konkreten Fragestellung aber ergebnisoffen. Die Bürger:innen bzw. die beteiligten Akteur:innen (oftmals u. a. Wissenschaftler:innen) stehen im Mittelpunkt und definieren gemeinsam das Ergebnis, oder kommen je nach Verfahren auch zu unterschiedlichen Ergebnissen. Mit Blick auf die Ergebnisse ist für gute Verfahren der Öffentlichkeitsbeteiligung das Erwartungsmanagement wichtig: Welche Form sollen die Ergebnisse haben und welchen inhaltlichen Spielraum haben die beteiligten Akteure? Für einen optimalen Austausch mit den Bürger:innen in der Wissenschaftskommunikation und in Beteiligungsverfahren braucht es daher eine kontiinuierliche Weiterentwicklung von Prozessen und Methoden innerhalb der Verfahren selbst. Der Beitrag gibt einen Überblick über verschiedene Formen des Feedbacks und beleuchtet konkrete Feedbackmethoden anhand von Praxisbeispielen.

E. Wollmann (✉)
ifok GmbH, Düsseldorf, Deutschland,
E-Mail: eva.wollmann@ifok.de

J. Birkenhäger
ifok GmbH, Berlin, Deutschland
E-Mail: jacob.birkenhaeger@ifok.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,*
https://doi.org/10.1007/978-3-658-39582-7_20

1 Einführung: Feedback in der Wissenschaftskommunikation und Öffentlichkeitsbeteiligung

Aktuelle und fortwährende Krisen wie die Corona-Pandemie oder der Klimawandel machen deutlich, wie bedeutend wissenschaftliche Fakten und ihr Verständnis für die Bewältigung gesellschaftlicher Herausforderungen und die Akzeptanz damit verbundener politischer Maßnahmen sind. In einer vielfältigen Wissensgesellschaft mit zahlreichen Informationsangeboten, in der jede:r ohne viel Aufwand die eigene Interpretation der Wirklichkeit verbreiten kann, wächst die Bedeutung von guter Kommunikation über Wissenschaft und Forschung. Im Kontext von Verfahren der Bürger:innen- und Öffentlichkeitsbeteiligung ist dabei die erfolgreiche Vermittlung wissenschaftlichen Wissens und neuester Forschungserkenntnisse als ein gemeinsamer Ausgangspunkt des Prozesses relevant. Diese Informationen müssen daher die Zielgruppe erreichen und die kommunizierten Inhalte den Beteiligten nachhaltig im Gedächtnis bleiben. Aus diesem Grund ist das Einholen von Feedback der Beteiligten bei Verfahren der Bürger:innen- und Öffentlichkeitsbeteiligung und damit verbundener Wissenschaftskommunikation ein wichtiger Bestandteil.

Die folgenden drei Beispiele machen deutlich, warum Feedback so wichtig für erfolgreiche Verfahren der Öffentlichkeitsbeteiligung ist. Der vorliegende Praxisbeitrag beschreibt im Anschluss das ‚Warum?' und ‚Wie?' von Feedback in der Bürger:innen- und Öffentlichkeitsbeteiligung als Teil der Verfahrensevaluation, und stellt konkrete Beispiele für Feedbackmethoden vor.

Einblicke in die Praxis: O-Töne

Scoping-Termin für ein wichtiges Infrastrukturvorhaben in einer nordrhein-westfälischen Großstadt. Nach einer ersten Bürger:innenkonferenz geht es nun darum, das erarbeitete Beteiligungskonzept unter Mitwirkung der Öffentlichkeit zu finalisieren. Als es um die Frage geht, wie das Vorhaben, wichtige Verfahrensschritte und Beteiligungsmöglichkeiten kommuniziert werden, meldet sich ein Schüler aus der Gegend zu Wort: „Wenn ihr die junge Generation erreichen wollt, dann dürft ihr nicht nur mit Postwurfsendungen und einem E-Mail-Newsletter arbeiten, sondern dann müsst ihr auf YouTube sein", so sein Credo. (ifok-Veranstaltung 2017. [Eigenes Gesprächsprotokoll])

***
Öffentlichkeitsbeteiligung zur Kulturpolitik eines Bundeslandes. Nach den ersten Workshops melden die Beteiligten zurück, dass die Themen zu breit und zu komplex seien, und es kaum möglich sei, konkrete Empfehlungen zu entwickeln. Man brauche mehr Austausch und Zeit zur Vertiefung der Themen. (ifok-Veranstaltung 2021. [Eigenes Gesprächsprotokoll])

***
EU-Bürger:innenforum zur Zukunft Europas. Nach einem ersten Zusammentreffen zufällig ausgewählter EU-Bürger:innen mit EU-Politiker:innen herrscht Ernüchterung: Die Bürger:innen beklagen, dass sie teilweise kaum zu Wort gekommen, und keine richtige Diskussion möglich gewesen sei. Stattdessen seien zusammenhanglose Statements der verschiedenen Politiker:innen aneinandergereiht worden, ohne Bezug aufeinander, und ohne Ergebnisse. (ifok-Veranstaltung 2022. [Eigenes Gesprächsprotokoll])

2 Warum und wofür? Feedback als Teil der Evaluation

Verfahren der Bürger:innen- und Öffentlichkeitsbeteiligung beziehen oftmals wissenschaftliches Wissen und konkrete Forschungserkenntnisse ein, sind im Rahmen der konkreten Fragestellung aber ergebnisoffen. Die Bürger:innen bzw. die beteiligten Akteur:innen (oftmals u. a. Wissenschaftler:innen) stehen im Mittelpunkt und definieren gemeinsam das Ergebnis, oder kommen je nach Verfahren auch zu unterschiedlichen Ergebnissen. Mit Blick auf die Ergebnisse ist für gute Verfahren der Öffentlichkeitsbeteiligung das Erwartungsmanagement wichtig: Welche Form sollen die Ergebnisse haben und welchen inhaltlichen Spielraum haben die beteiligten Akteur:innen? Geht es zum Beispiel darum, politische Empfehlungen, Gesetzesentwürfe, Rückmeldung zu möglichen Trassenverläufen einer neuen Bahnstrecke oder räumliche Ideen für die Planung der Neugestaltung einer Innenstadt zu erarbeiten?

Ergebnisoffene Verfahren sind lernende Verfahren, die methodisch und strukturell auf die Bedürfnisse der beteiligten Akteur:innen reagieren müssen. Für den Erfolg des Verfahrens ist es zentral, dass die beteiligten Akteur:innen

mit der Durchführung zufrieden sind, um hinterher auch das Ergebnis als legitim zu empfinden (auch wenn manche Akteur:innen den Ergebnissen final nicht zustimmen können oder wollen). Um diese Verfahrenslegitimation zu erreichen, sollten Rückmeldungen und Kritik der Beteiligten gehört und, wenn sinnvoll und möglich, auch berücksichtigt werden. Wichtig ist aber: Die konzeptionelle und methodische Steuerung des Verfahrens liegt nicht bei den beteiligten Akteur:innen, sondern sollte von erfahrenen Gestalter:innen und Moderator:innen der Öffentlichkeits- und Bürger:innenbeteiligung umgesetzt werden. Andernfalls wird der Prozess schnell unstrukturiert und beliebig.

2.1 Funktionen von Feedback

Das Feedback in der Wissenschaftskommunikation und Öffentlichkeitsbeteiligung kann grundsätzlich zwei Funktionen haben, die je unterschiedliche Evaluationsansätze haben. Einerseits kann Feedback dazu beitragen, ein laufendes Verfahren zu verbessern. Es handelt sich dann um einen Beitrag zu einer formativen bzw. prozessualen Evaluation. Andererseits soll Feedback helfen, aus abgeschlossenen Verfahren für zukünftige Projekte zu lernen. Entsprechendes Feedback zahlt auf die summative Evaluation ein (vgl. dazu auch den Beitrag von Sophia Volk in diesem Band).

Im Rahmen der formativen Evaluation mit Fokus auf den Prozess wird laufend oder zu bestimmten Zeitpunkten der Kommunikations- und Beteiligungsmaßnahmen Feedback der Zielgruppe(n) eingeholt, um daraus für die weitere Verfahrensgestaltung zu lernen und diese anzupassen (vgl. Volk in diesem Band). Grundsätzlich ergeben Feedback und eine formative bzw. prozessuale Evaluation dann Sinn, wenn ein Verfahren über einen längeren Zeitraum läuft und Verbesserungen überhaupt noch implementiert werden können. Dies kann zum Beispiel im Rahmen eines Bürger:innenrates erfolgen, der über einen bestimmten Zeitraum mehrfach tagt und dessen Sitzungen verbessert werden können. Auch wenn im Rahmen eines Beteiligungsprojekts mehrfach das gleiche Format mit verschiedenen Teilnehmenden umgesetzt werden soll, beispielsweise eine Bürger:innenkonferenz oder Fokusgruppe, empfiehlt sich dieses Vorgehen, sofern die exakt gleiche Umsetzung nicht aus anderen (verfahrenstechnischen oder wissenschaftlichen) Gründen notwendig ist. Nicht sinnvoll und umsetzbar ist eine formative Evaluation bei einmaligen Veranstaltungen oder Kommunikationsmaßnahmen.

Die summative Evaluation wird zum Ende eines Projekts gemacht und bezieht Erkenntnisse aus dem ganzen Verfahrensverlauf ein (vgl. Volk in diesem

Band). So lassen sich die Ergebnisse feststellen und Schlüsse für zukünftige, ähnliche Prozesse ziehen; eine Wirkung auf das evaluierte Verfahren oder die abgeschlossenen Kommunikationsmaßnahmen selbst ist aber nicht möglich. Denn das Feedback der beteiligten Akteur:innen holt man im Rahmen der summativen Evaluation erst am Ende des Prozesses ein, zum Beispiel im Rahmen eines Fragebogens oder einer Feedbackwand (siehe Beschreibung der Methoden im zweiten Teil des Beitrags).

2.2 Inhalte von Feedbacks und was es zu beachten gilt

Inhalt von Evaluationen können verschiedene Aspekte von Verfahren der Öffentlichkeits- und Bürger:innenbeteiligung sein. Das beginnt bei der Vorbereitung, bei der Einladung zum Verfahren und der Begleitung der Teilnehmenden über die konkrete Veranstaltungs- und Prozessorganisation, die methodische, technische und organisatorische Umsetzung und Moderation, die Neutralität, Ausgewogenheit und gleichberechtigte Einbeziehung aller Beteiligten, die Transparenz bis hin zur Qualität und Umsetzung der Ergebnisse. Zu allen Bereichen lässt sich Feedback der Beteiligten einholen, sowohl im Rahmen einer summativen als auch – je nach Prozessschritt für bestimmte Bereiche – im Rahmen einer formativen Evaluation.

Wie Feedback konkret erhoben werden kann, folgt im zweiten Teil des Beitrags. Grundsätzlich richten sich die eingesetzten Methoden nach Zeitpunkt und Zweck des Feedbacks, aber auch nach der Gruppengröße, dem Verfahrensaufbau und dem Grad der Formalität sowie danach, ob das Verfahren online oder analog stattfindet. Für jede Form des Feedbacks gilt es, folgende Grundregeln zu beachten:

- Das Feedback sollte *fester Bestandteil des Verfahrens* sein. Es muss *für alle Beteiligten zugänglich* sein und alle müssen gleichermaßen gehört werden.
- Feedback sollte *transparent sein* und transparent umgesetzt werden, mindestens für die beteiligten Personen.
- *Bei formativer Evaluation: Die mit der Durchführung betrauten Personen müssen auf das Feedback reagieren.* Nicht jedes Feedback muss umgesetzt werden, aber: Wo möglich und sinnvoll, sollten Anpassungen vorgenommen werden und in jedem Fall die Reaktion mindestens innerhalb des Verfahrens transparent und nachvollziehbar gemacht werden.

Wie Feedback in der Durchführung von Öffentlichkeits- und Bürger:innenbeteiligung wirken kann, zeigen die Beispiele vom Anfang: Im Rahmen eines Infrastrukturvorhabens einer nordrhein-westfälischen Großstadt haben die Vorhabensträger neue Wege der Kommunikation auch über WhatsApp und Facebook ausprobiert, sowie YouTube-Videos erstellt. Die zuständige Ministerin hat den Beteiligungsprozess zur Kulturpolitik ausgedehnt und den Akteur:innen mehr Raum für die Arbeit an ihren Ideen gegeben. In beiden Fällen hat die Rückmeldung also zur Verbesserung der Verfahren beigetragen. Und auf europäischer Ebene überlegen die durchführenden Institutionen jetzt, wie sie bei den nächsten Zusammentreffen von Bürger:innen und Politiker:innen in Arbeitsgruppen und gemeinsamen Sitzungen die Kommunikation und Zusammenarbeit verbessern können.

3 Praxisbeispiele kreativer Feedbackmethoden und -formate

Bevor einzelne Methoden vorgestellt werden, geht es um die Frage, wozu es eigentlich spezifische Methoden zum Einholen von Feedback braucht. Methoden sind das Werkzeug und die Art und Weise, wie ein Ziel, in dem Fall das konstruktive Feedback, erlangt werden kann. Sie helfen dabei, die unterschiedlichen Bedürfnisse der Teilnehmenden systematisch zu erfassen sowie laute und leise Stimmen in Workshops und Veranstaltungen gleichwertig zu Wort kommen zu lassen.

Doch auch wenn innovative Methoden bei der Konzeption jener Verfahren eine große Rolle spielen, sei an dieser Stelle ausdrücklich gesagt: Es kommt nicht immer darauf an, innovativ und neu zu denken, vielmehr steckt der Teufel auch bei altbewährten Methoden im Detail. Entscheidend ist, den Kern der Methode zu verstehen und sie dementsprechend umzusetzen, um das gewünschte Feedback zu generieren. Auch die Feedback-Fragen sollten im Vorfeld gut und sorgfältig durchdacht sein. Eine Anregung für die Wahl der geeigneten Methoden können folgende Fragen liefern:

- Was ist der Nutzen der Feedbackabfrage? Was soll mit dem Feedback erreicht werden?
- Welche Art des Feedbacks wird benötigt? Geht es darum, ein laufendes Verfahren zu verbessern (formative Evaluation), oder soll ein abgeschlossenes Projekt bewertet werden (summative Evaluation)?
- Geht es darum, inhaltliches oder prozessuales Feedback einzuholen?

- Zu welchem Zeitpunkt im Verfahren/Projekt soll und kann Feedback eingeholt werden?
- Welches Budget steht dafür zur Verfügung?

Je nachdem, wie die Fragen beantwortet werden, lässt sich das Design des Feedbacks anlegen und eine konkrete Methode auswählen. Beispielsweise kann das Feedback entweder direkt nach einer Veranstaltung oder im Nachgang – einige Tage oder wenige Wochen später – eingeholt werden. Auch eine Kombination der beiden Varianten ist möglich.

Die folgenden Ausführungen geben eine Übersicht über einige ausgewählte Feedbackmethoden und ermöglichen einen Einblick in Verfahren, in denen umfangreiche Feedbackprozesse Einzug gefunden haben.

3.1 Das Blitzlicht

Die Blitzlicht-Methode eignet sich besonders dazu, Meinungs- und Stimmungsbilder einer mittelgroßen Gruppe von bis zu 20 Teilnehmenden zum Ausdruck zu bringen. Die Moderation stellt eine gezielte Frage zum Prozess oder einer Veranstaltung, zu der sich Freiwillige zu Wort melden. Dabei dürfen die restlichen Teilnehmenden das Gesagte nicht kommentieren und auch die Moderation leitet keine Diskussion zu dem Gesagten an. Lediglich Verständnisfragen durch die Moderation sind zugelassen. Das aus einer Blitzlichtrunde entstandene Bild kann helfen, die Arbeitssituation positiv zu gestalten und lösungsorientiert zu verändern. Für diese Feedbackmethode sollte ausreichend Zeit eingeräumt werden. In der Praxis hat sich die Regel 1–2 min pro Person bewährt.

Eine gute Visualisierung und Ergebnissicherung durch die Moderation spielen eine besondere Rolle bei der Durchführung des Blitzlichts. Sie soll den Teilnehmenden Wertschätzung und Verständnis vermitteln und Raum für Bedürfnisse und Befindlichkeiten einräumen. Jede Aussage, egal ob zu den Inhalten des Projekts, dem eigenen Wohlbefinden oder zum Wetter, wird aufgenommen. Dabei entscheidet die Moderation nicht über die Wichtigkeit einer Aussage! Mit etwas Abstand lassen sich auch aus scheinbar unwichtigen Bemerkungen wertvolle Hinweise gewinnen.

Die Blitzlichtrunde wird häufig am Ende einer Veranstaltung oder eines Prozesses durch eine konkrete Frage eingeleitet. (z. B. „Was nehme ich mit?“; „Was ist mir klar geworden?“; „Was ist für mich unklar geblieben?“). Sie kann jedoch zu jeder Zeit und flexibel eingesetzt werden. Am Ende der Runde sollte die Moderation die Anmerkungen bündeln und eventuell Schlussfolgerungen zur

Diskussion stellen. Vor Ort werden für diese Methode lediglich Moderationswände und -karten sowie Stifte benötigt. Im digitalen Raum eignen sich virtuelle Whiteboards wie bspw. Mural, Miro, Conceptboard etc.

Varianten
Die Teilnehmenden finden sich kurz in spontanen Zweier- oder Dreiergruppen zusammen und diskutieren über eine durch die Moderation vorgegebene Fragestellung. Im Plenum werden die Zweier- oder Dreierteams um ein kurzes Statement gebeten. Die Zahl der Blitzlichter kann dadurch wesentlich niedriger ausfallen (bspw. max. fünf). Bei Bedarf können auch Gegenstände genutzt werden, die der Reihe nach herumgegeben werden. Dabei haben die Teilnehmenden bspw. die Wahl zwischen einem Koffer (alternativ einer Tasche) und einem Radiergummi/Blitz. Der Koffer steht für die wichtigsten „Takeaways“ aus dem Prozess, wohingegen der Radiergummi/Blitz für die Dinge steht, die sich die Teilnehmenden anders gewünscht hätten. Beide Varianten lassen sich sowohl vor Ort als auch digital umsetzen.

3.2 Feedbacksitzung

Mithilfe einer dem Prozess/Verfahren nachgelagerten Feedbacksitzung lässt sich in einer offenen Runde konstruktives Feedback, sei es Lob oder auch Kritik der Teilnehmenden, einholen. Eine solche Sitzung lässt sich ebenfalls direkt nach bzw. zum Ende einer Veranstaltung umsetzen. Erfahrungsgemäß eignet sich hierbei die Kopplung weiterer Methoden, wie beispielsweise die Arbeit mit einer Feedbackwand (siehe Abschn. 3.3).

Im Bürger:innenrat Klima 2021 bspw. haben unter der Schirmherrschaft von Bundespräsident a. D. Horst Köhler insgesamt 160 zufällig ausgewählte Bürger:innen Empfehlungen und Leitsätze erarbeitet, wie Deutschlands Klimapolitik gestaltet werden müsste, um die Pariser Klimaschutzziele einzuhalten. Ein insgesamt sehr komplexer und umfangreicher Prozess, in den neben Bürger:innen auch Akteur:innen (Expert:innen) aus Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft eingebunden waren. In insgesamt zwölf Sitzungen haben die Bürger:innen in der Auseinandersetzung mit den Informationen seitens der Expert:innen in intensiver Kleingruppenarbeit konkrete Maßnahmenbündel erarbeitet und im Anschluss an die Politik übergeben. Um diesen komplexen Prozess zu evaluieren, hat sich das Projektteam für eine Kombination aus formativer und summativer Evaluation entschieden. Diese sah wie folgt aus:

Zur Halbzeit des Bürger:innenrates hat eine „Feedbacksitzung" mit zufällig ausgewählten Teilnehmenden (insgesamt acht Bürger:innen der 160 Teilnehmenden des Bürger:innenrates, jeweils zwei Teilnehmende pro Themenfeld des Bürge:innnenrates) stattgefunden.

Die folgende Übersicht zeigt einen exemplarischen Ablauf der Feedbacksitzung:

1. Kurze Begrüßung und Dank, Vorstellung der Anwesenden, Intention und Erwartungsmanagement, kurze Runde zum Ankommen (nur Teilnehmende): Welches Gefühl ist bei Ihnen gerade am stärksten, wenn Sie an den Bürger:innenrat denken? Bitte nur einen Satz.
2. Fokus auf konkrete Fragen an die Teilnehmenden, dann immer einmal die Runde durch, sodass jede:r etwas sagen kann. Folgende Themen wurden diskutiert: Arbeitsweise und Struktur, Referent:innen, Qualität der Moderation und Diskussion, Technik und Informationsmaterial, Transparenz des Bürger:innenrat-Prozesses, Klarheit der Zielsetzung (Ist nachvollziehbar, worum es beim Bürger:innenrat Klima geht und warum er einberufen wurde?)
3. Kurze Abschlussrunde

Die Sitzung hat dem Projektteam geholfen, den weiteren Prozess zu optimieren und an die Bedürfnisse der Beteiligten anzupassen. Beispielsweise forderten die Bürger:innen insgesamt mehr Zeit für Austausch innerhalb der Kleingruppen, wodurch die Bürger:innen in den nachfolgenden Sitzungen mehr Zeit zur Diskussion und Reflexion bekamen.

Um den Gesamtprozess des Bürger:innenrates am Ende umfangreich auszuwerten, wurden zeitnah nach dessen Abschluss alle am Prozess beteiligten Akteur:innen um Feedback mithilfe eines Fragebogens (siehe auch Böhmert und Abacioglu in diesem Band) gebeten. Dieser wurde den Teilnehmenden digital mit einer Rücklauffrist von 14 Tagen zur Verfügung gestellt und der Link zur Teilnahme per per E-Mail versendet. Damit die Teilnehmenden motiviert und mit einem guten Erinnerungsvermögen Rückmeldung zum Prozess geben können, ist die zeitliche Komponente besonders wichtig. Aus diesem Grund hat die Moderation während der Veranstaltungen des Verfahrens bereits den Fragebogen angekündigt und die Teilnehmenden des Bürger:innenrates zur Teilnahme und aktiven Mitarbeit motiviert. Von insgesamt 160 Teilnehmenden des Bürger:innenrates haben sich 148 Teilnehmende an dem Fragebogen beteiligt, 109 Teilnehmende haben ihn abgeschlossen und ihre Eingaben abgeschickt. Als zentrale Erkenntnis aus den Rückmeldungen lässt sich festhalten, dass sich die Teilnehmenden solcher Prozesse mehr Zeit zur Deliberation im Prozess

gewünscht hätten sowie die Verstetigung solcher und ähnlicher Formate. Dennoch: Mit 94,6 % würde der Großteil mit großer Wahrscheinlichkeit wieder an einem Bürger:innenrat teilnehmen. Die Stimmung wurde von den Teilnehmenden als gut (56,8 %) bis sehr gut (31,2 %) empfunden, und neben dem Gemeinschaftsgefühl wurde vor allem die Diskussion in den Kleingruppen besonders hervorgehoben. Ausschlaggebender Grund für die Teilnahme war das Interesse am Thema und der Wunsch zur Mitwirkung. In der Auswertung wurde deutlich, dass die Veranstaltung insgesamt besser bei Menschen mit hohem Bildungsabschluss sowie bei Frauen ankam.

3.3 Feedbackwand

Die Feedbackwand ist ein gängiges Medium, welches es Teilnehmenden ermöglicht, ihr Feedback (Wünsche, Anregungen sowie sonstige Rückmeldungen und Fragen) für alle sichtbar im (virtuellen) Raum anzubringen. Hierfür steht den Teilnehmenden eine Wand (Flipchart vor Ort oder digitales Whiteboard online) zur Verfügung. Da am Ende eines Prozesses/einer Veranstaltung häufig die Motivation der Teilnehmenden sinkt, empfiehlt es sich, explizit Zeit für eine kurze und stille Reflexionsrunde einzubauen und den Teilnehmenden konkrete Frage- oder Aufgabenstellungen für Feedback mitzugeben. Dies kann beispielsweise durch Icons und einer unterschiedlichen Farbgebung unterstützt werden. Vor Ort werden dafür Moderationswände, verschiedenfarbige Klebepunkte, Stifte und Brownpaper benötigt. Im digitalen Raum kann mit virtuellen Whiteboards, Symbolen und Skalenabfragen gearbeitet werden. Hier empfiehlt sich, den Teilnehmenden eine technische Einführung in die Nutzung des Tools zu geben.

Hier ein Beispiel für eine Feedbackwand (Abb. 1):

Varianten

Für die Ausgestaltung der Feedbackwand gibt es verschiedene Varianten.

Digitales Whiteboard: Im Rahmen des Strategiedialogs Automobilwirtschaft Baden-Württemberg (SDA) hat 2021 das Bürger:innenforum „Digitalisierung der



Abb. 1 Feedbackwand. (Quelle: eigene Darstellung)

Mobilität“ als Online-Veranstaltung stattgefunden. Ziel des Bürger:innenforums war es, die Nutzer:innenperspektive als einen Grundpfeiler digitaler Geschäftsmodellentwicklungen in den aktuellen Wandel der Automobilindustrie einfließen zu lassen. In insgesamt vier digitalen Sitzungen haben die Teilnehmenden über das Thema diskutiert und Handlungsempfehlungen an Politik, Wirtschaft und Gesellschaft entwickelt. Für eine kontinuierliche und transparente Dokumentation wurden die Diskussionen und Ergebnisse an einem umfangreichen digitalen Whiteboard festgehalten. Die Teilnehmenden erhielten die Möglichkeit, sich zu jeder Zeit frei auf dem Whiteboard zu bewegen und die Inhalte der Sitzungen einzusehen, etwa Präsentationen und Informationen zu Expert:innen und dem Prozess insgesamt, sowie Feedback, Fragen, Anregungen und Kritik zum Prozess zu geben. Diese Variante ermöglichte den Teilnehmenden, anonymes Feedback über den gesamten Zeitraum des Prozesses zu geben, auf welches noch während des Prozesses eingegangen werden konnte.

Insgesamt hilft diese transparente und kontinuierliche Art des Feedbacks, den Prozess noch während der Laufzeit zu optimieren und auf die unterschiedlichen Bedürfnisse der Teilnehmenden zu reagieren. Ein positiver Nebeneffekt: Die Teilnehmenden fühlen sich gehört. Im besten Fall sorgt ein kontinuierlicher Blick auf die Feedbackwand zu Beginn jeder Sitzung für eine positiven Arbeitsatmosphäre – denn: Die Teilnehmenden fühlen sich in den Prozess integriert und wertgeschätzt.

Hier ein Einblick in mögliche Fragen des digitalen Whiteboards mit integrierter Feedbackwand:

- Was war super/hat gut funktioniert und sollte beibehalten werden?
- Was war nicht so gut und sollten wir ändern?
- Gibt es Ideen, um noch effektiver im digitalen Whiteboard arbeiten zu können?
- Allgemeine Gedanken und Fragen

Die Punkteabfrage: Mithilfe dieser Variante können die Teilnehmenden konkrete Fragen zum Prozess/der Veranstaltung auf Moderationswänden „bepunkten“ und dadurch bewerten. Hierbei können mit Stimmungsbarometern wertvolle Erkenntnisse über Bedürfnisse und weitere steuerungsrelevante Informationen gewonnen werden. Dazu bekommen alle Teilnehmenden dieselbe Anzahl an Punkten (analog: Klebepunkte – digital: Symbole). Die Moderation leitet die Fragestellungen für das Feedback ein und gibt den Teilnehmenden einen Moment, sich auf den jeweiligen Skalen/Stimmungsbarometern zu verorten. Die Teilnehmenden dürfen ihre Punkte auf dem für sie relevantesten Thema häufen,

oder aber sie verteilen die Punkte auf die unterschiedlichen Kategorien. Auch eine Kombination aus beidem ist möglich. Nachdem sie ihre Punkte vergeben haben, fasst die Moderation die Ergebnisse zusammen und gibt die Rangfolge der einzelnen Kategorien und Bewertungen wieder. Je nach Fragestellung sollte darauf geachtet werden, ob die Punkteabfrage in einem geschützten Rahmen stattfindet und die Teilnehmenden bei der Abfrage anonym bleiben, oder ob die Abfrage offen und für alle transparent verläuft.

Die Skalenabfrage: Mithilfe vorgefertigter Skalen, z. B. von 0 bis 10 oder von unwichtig bis sehr wichtig, lassen sich sehr unterschiedliche Inhalte abfragen. Zudem können sie während eines Prozesses oder am Ende für Feedback und Evaluation eingesetzt werden.

3.4 Aufstellung im Raum

Die Aufstellung im Raum ermöglicht es nicht nur Informationen und Stimmungsbilder für alle sichtbar zu machen, sondern bringt die Teilnehmenden in Bewegung. Die Methode kann also auch als Energizer genutzt werden. Hierzu gibt die Moderation Fragen vor, nach denen sich die Teilnehmenden im Raum positionieren. Zu Beginn eignet sich eine einfache Frage als Ice-Breaker wie beispielsweise *„Wie geht es dir?“*. Im Raum wird ein Punkt als „sehr gut“ und ein anderer, durch eine imaginäre Linie mit dem ersten verbunden, als „ausgelaugt und müde“ (o. ä.) definiert. Die Teilnehmenden stellen sich entlang der imaginären Linie auf, und einzelne können je nach Zeit von der Moderation auch zu ihrer Positionierung befragt werden. Anschließend kann mit den Themen und Fragen fortgefahren werden, die auf die konkrete Evaluation der Veranstaltung oder des Prozesses einzahlen.

Ein weiterer schöner Nebeneffekt: Die Methode sorgt für ein Gemeinschaftsgefühl und regt die Kommunikation der Teilnehmenden untereinander an. Für diese Methode werden vor Ort lediglich die entsprechenden Räumlichkeiten benötigt. Für eine digitale Alternative lässt sich die Aufstellung im Raum mithilfe eines vorbereiteten Whiteboards umsetzen. Dazu schreiben die Teilnehmenden ihre Namen auf ein Kärtchen und platzieren es an der entsprechenden Stelle im Whiteboard.

Varianten

Während der Veranstaltung geht ein Mitglied des Organisationsteams durch den Raum und sammelt O-Töne und Statements der Teilnehmenden. Diese werden für alle gut sichtbar auf Moderationswände notiert und an unterschiedlichen Orten im

Raum platziert. In einer abschließenden Runde verliest die Moderation die verschiedenen Statements und bittet die Teilnehmenden, sich an dem jeweiligen Statement zu positionieren, das ihrer Haltung/Meinung am ehesten entspricht. Als Ergänzung können die Teilnehmenden dann zum von ihnen ausgewählten Statement interviewt werden.

4 Fazit

Gute Wissenschaftskommunikation und Beteiligung gelingen nur dann, wenn die Zielpersonen und Beteiligten aktiv an den Prozessen teilnehmen und sich mit ihren Ideen und Anregungen einbringen. Für einen optimalen Austausch mit Bürger:innen in der Wissenschaftskommunikation und in Beteiligungsverfahren braucht es eine kontinuierliche Weiterentwicklung von Prozessen und Methoden. Feedback ist hierfür ein essenzieller Bestandteil.

Nachdem die Funktionen und Abläufe verschiedener Feedbackmethoden mit entsprechenden Praxisbeiträgen näher beleuchtet wurden, schließt der Beitrag mit folgendem Appell:

1. Ergebnisoffene Verfahren sind lernende Verfahren und deshalb auf Feedback während des Prozesses angewiesen. Trotzdem sollte Feedback auch am Ende des Prozesses nicht außer Acht gelassen werden.
2. Feedback lässt sich bereits niedrigschwellig im Verfahren integrieren. Methoden wie die Aufstellung im Raum ermöglichen die Teilnahme aller und sorgen nicht nur für Rückmeldung zum Prozess, sondern wirken gleichzeitig als aktivierendes Element während laufender Veranstaltungen.
3. Die ausgewählten Feedbackmethoden müssen nicht immer innovativ sein. Auch klassische Methoden können Interessantes zutage fördern. Je konkreter die Fragen für die Formulierung von Feedback, desto präziser und konstruktiver fällt das Feedback der Teilnehmenden aus.
4. Aus diesem Grund empfiehlt es sich, sich bereits während der Konzeption des Verfahrens konstruktiv und kritisch mit dem Prozess auseinanderzusetzen und Feedback von Anfang an einzuplanen. Gut durchdachtes, strukturiertes und frühzeitiges Feedback trägt zum Erfolg von Austausch und Beteiligung bei, da es ermöglicht zu lernen und auf die Bedürfnisse der beteiligten Akteur:innen in einem Prozess zu reagieren. Aus diesem Grund ist Feedback zum Prozess gekoppelt mit Aspekten einer summativen Evaluation oftmals eine gute Lösung.

5. Zu guter Letzt braucht gutes Feedback eine Offenheit im Prozess. Feedback ist erst dann legitim, wenn es Veränderungen ermöglicht und zulässt.

Eva Wollmann ist seit 2019 Mitglied des ifok-Teams und seit 2020 als Consultant in den Themenfeldern Mobilität und Kommunikation & Kampagnen tätig. Ihre inhaltlichen Schwerpunkte liegen in der Methodenentwicklung, Workshop- und Veranstaltungskonzeption, Durchführung und Nachbereitung von analogen sowie virtuellen Dialogprozessen.

Jacob Birkenhäger ist Geschäftsfeldleiter für Deliberation, Open Government und Demokratie. Er betreut bei ifok seit 2015 verschiedenste Dialogprozesse und beschäftigt sich mit deliberativen Ansätzen der politischen und gesellschaftlichen Steuerung, Open Government-Ansätzen und Strukturen für eine zukunftsfähige Demokratie. Als Projektleiter verantwortete er EU-, bundes- und landesweite Dialogprozesse, u. a. die ersten bundesweiten Bürger:innenräte zur Demokratie und zu Deutschlands Rolle in der Welt sowie die EU-Bürger:innenforen im Rahmen der Konferenz zur Zukunft Europas.



Praxisbeitrag: Multimethodenansatz in der Evaluation am Beispiel der Dialogveranstaltung „Mensch Wissenschaft!“

Markus Gabriel, Isabella Kessel, Thomas Quast und Eva Roth

Zusammenfassung

Mit „Mensch Wissenschaft!“ hat die Robert Bosch Stiftung in einem mehrstufigen und durch eine Evaluation begleiteten Verfahren ein dialogisches und partizipatives Angebot der Wissenschaftskommunikation entwickelt. Dabei handelt es sich um eine zweitägige Veranstaltung mit dem Ziel, einen gleichberechtigten Dialog zwischen Bürger:innen und Wissenschaftler:innen zu initiieren und auf diese Weise gegenseitiges Verständnis und Vertrauen zu fördern. Das Angebot ist als Blaupause für Dialogformate zwischen Bürger:innen und Wissenschaftler:innen gedacht. Die begleitende formative Evaluation ist ein praxisorientiertes Beispiel für den Einsatz verschiedener aufeinander abgestimmter und verschränkter (quantitativer und qualitativer)

M. Gabriel · T. Quast (✉)
com.X Institut, Bochum, Deutschland,
E-Mail: thomas.quast@comx-forschung.de

M. Gabriel
E-Mail: markus.gabriel@comx-forschung.de

I. Kessel
Translake, Konstanz, Deutschland,
E-Mail: isabella.kessel@translake.org

E. Roth
Robert Bosch Stiftung GmbH, Stuttgart, Deutschland,
E-Mail: eva.roth@bosch-stiftung.de

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,*
https://doi.org/10.1007/978-3-658-39582-7_21

Methoden in einem Evaluationsvorhaben, auch im engeren Sinne eines Multimethoden-Forschungsansatzes- oder Mixed-Method-Designs. Der Beitrag gibt zudem eine kurze theoretische Einführung in das Thema multimethodischer bzw. Mixed-Method-Designs.

1 Multimethodenansätze/Mixed Methods in der Evaluation von Wissenschaftskommunikation

Bevor anhand der Evaluation des Dialogformats „Mensch Wissenschaft!" der Robert Bosch Stiftung ein Beispiel für einen Multimethodenansatz in der Praxis vorgestellt wird, findet sich hier eine kurze theoretische Einführung in das Thema Multimethodenan- bzw. -einsatz.

Für Evaluationen des com.X Institus, und sicher in der Evaluationspraxis generell, ist der Einsatz verschiedener Methoden in einem Evaluationsvorhaben sehr üblich, auch in der Evaluation von Formaten der Wissenschaftskommunikation. Das ergibt sich schon – und zwar nicht nur bei Großformaten der Wissenschaftskommunikation wie etwa den seit 2000 vom BMBF[1] und WiD[2] ausgerichteten Wissenschaftsjahren – aus der Notwendigkeit, in einer umfassenderen Evaluation etwa eine Vielzahl von Angeboten, Formaten, Ziel- und Akteursgruppen sowie Perspektiven berücksichtigen zu müssen (Gabriel und Quast 2006; Gabriel und Warthun 2017): Z. B. Präsenzausstellungen und -diskussionen, virtuelle Angebote, on- und offline (Begleit-)Kommunikation/ Information bzw. Kinder- und Jugendliche, die „allgemeine" Öffentlichkeit, Wissenschaft oder Politik. Als geeignete Methoden zur Evaluation der genannten Beispiele bieten sich ad hoc an: (Teilnehmende) Beobachtung, quantitative und qualitative Befragung bzw. Leitfadeninterviews-/gespräche, Nutzungsdatenanalyse sowie Inhalts- bzw. Medienanalysen digitaler und nicht digitaler Medien (sowohl von owned media wie earned media)[3]. Dazu könnte auch eine Wirtschaftlichkeitsanalyse kommen, um die Effizienz des Vorhabens auch aus dieser

[1] Bundesministerium für Bildung und Forschung.

[2] Wissenschaft im Dialog gGmbH.

[3] „Owned Media" umfasst alle eigenen Medien und Kanäle, über die hinsichtlich Inhalt und Gestaltung die Steuerungs- und Kommunikationshoheit besteht. „Earned Media" bezeichnet die „verdiente" Sichtbarkeit, indem Inhalte von anderen Akteur:innen (Medien,

Perspektive zu betrachten. Alle eingesetzten Methoden könnten aber unabhängig voneinander bzw. nur verbunden durch den analytischen Blick auf das Gesamtformat zum Einsatz kommen und nur einen jeweils eigenen Bereich des Gesamtformats in den Blick nehmen. D. h., dass eine Vielzahl eingesetzter Methoden in *einem* Forschungs- oder Evaluationsvorhaben noch kein Multimethoden-Forschungsansatz oder Mixed- Method-Design im engeren Sinne sein muss.[4]

Unter multimethodischem bzw. Mixed-Method-Design ist, zumeist bzw. je nach wissenschaftlicher Diktion auch immer der Mix quantitativer und qualitativer Methoden in Bezug auf *einen* (Sub-)Gegenstand der Evaluation zu verstehen (Leech und Onwuegbuzie 2009; Tashakkori und Teddlie 1998).[5] Das ist z. B. der Fall, wenn Teilnehmende einer Veranstaltung der Wissenschaftskommunikation in der Breite *quantitativ* befragt und (zumeist selektiv) *qualitativ* interviewt werden, ggf. noch ergänzt um eine (teilnehmende) Beobachtung der Veranstaltung und alle Methoden zumindest in Teilen und aufeinander abgestimmt die gleiche Fragestellung anvisieren. In diesem Zusammenhang spricht man im engeren Sinne vom vielleicht bekanntesten Ziel[6] eines multimethodischen Designs, der Triangulation, die einerseits sinnvoll ist, um Methodeneffekte zu kontrollieren und spezifische Stärken verschiedener Methoden zu nutzen, auch um insgesamt eine bessere Erklärungsstärke, also im Prinzip Validität zu erzielen (Burke et al. 2007; Creswell und Plano-Clark 2007). Bei verschiedenen Erkenntnisinteressen lassen sich diese zudem jeweils oft besser, vertiefter und ggf. sogar ausschließlich nur durch eine spezifische Methode klären, d. h. hier haben wir eher sich ergänzende Methoden. Zudem

---

Zielgruppen etc.) unabhängig und aus eigenem Antrieb heraus erstellt bzw. weiterverbreitet werden. Der Begriff „Paid Media“ umschreibt alle Formen von kostenpflichtigen Marketing- und Kommunikationsmaßnahmen, um die Reichweite zu optimieren (Beilharz et al. 2020).

[4] Nach Tashakkori und Teddlie (2003) wären Mixed Method Designs oder Multimethod Research quasi als Untergruppen von Multiple Method Designs zu sehen.

[5] Auf den langen, teils bis in die 1930er Jahre zurückreichende Anwendung multimethodischer, quantitative und qualitative Methoden mixender Designs in der Sozialforschung, weist etwa Kelle (2015) in seiner knappen, aber profunden Erläuterung zum Mixed Methods Ansatz hin. Wobei die moderne Debatte zum Thema gegen Mitte bis Ende der 1990er Jahre an Fahrt aufnahm, weshalb auch viele moderne grundlegende theoretische Auseinandersetzungen mit dem Mixed Method Ansatz in den späten 1990ern und 2000er Jahre entstanden.

[6] „The most common and well-known approach to mixing methods is the Triangulation Design“ (Creswell und Plano-Clark 2007).

können – je nach zeitlicher Verschaltung der Methoden – auch Erkenntnisse, die eine qualitative Methode bereits geliefert hat, zur Präzisierung der Konstruktion (im Sinne einer Exploration) der zeitlich folgenden Methode dienen oder umgekehrt eine nachgeschaltete qualitative Methode zur Interpretation und Klärung qualitativer Ergebnisse herangezogen werden (Plano-Clark et al. 2008). In der Evaluationspraxis sind oftmals aber auch Überlegungen hinsichtlich mehrerer der beschriebenen und weiterer Effekte[7] multimethodischer Designs ausschlaggebend, um einen solchen Ansatz zu verfolgen.[8]

2 „Mensch Wissenschaft!" als Blaupause für Dialogformate zwischen Bürger:innen und Wissenschaftler:innen

2.1 Veranstaltungsidee und Ziele

In der Wissenschaftskommunikation wird zunehmend wichtiger, dass Wissenschaft nicht nur Erkenntnisse vermittelt – Stichwort „kommunikative Einbahnstraße", sondern sich in einem umfassenderen Sinne der Gesellschaft öffnet, in einen gleichberechtigten Austausch mit ihr tritt und den Dialog in

[7] „The four major types of mixed methods designs are the Triangulation Design, the Embedded Design, the Explanatory Design, and the Exploratory Design" (Creswell und Plano-Clark 2007). Der Typus der Einbettung (Embedded) ist unseres Erachtens etwa bei der Integration qualitativer Elemente in quantitative Befragungen (z. B. offene Fragen mit der Möglichkeit freier nicht vorgegebener Antwortmöglichkeiten) und umgekehrt noch nicht unbedingt als Methodenmix zu verstehen.

[8] In den letzten Jahren wurden auch Mixed-Mode-Designs diskutiert, die nicht mit Mixed Methods zu verwechseln sind. Hierbei geht es um unterschiedliche Zugangswege zu den Probanden bei Befragungen. Bei etwa bevölkerungsweiten Telefonbefragungen mit zufallsgenerierten Telefonnummern wird die Kombination von Festnetz- und Mobilanrufen (da viele Personen keine Festnetznummern mehr nutzen) oft schon als Mixed Mode bezeichnet. Dabei bleibt der Zugangskanal ja das Telefon, was methodisch kaum Implikationen hat. Interessanter wird es beim Mix von Online- und Telefonbefragung. Hier sind die Überlegungen, Non-Contact- und Non-Response-Raten insgesamt und in unterschiedlichen Subgruppen zu reduzieren. Unterschiedlichen Nutzerstrukturen bei klassischer Festnetz- sowie Mobil-Telefonie und auch des Internets sollen so ebenso Berücksichtigung finden, wie verändertes Nutzungsverhalten generell (Krug et al. 2014). Hier jedoch wird der methodenrelevante Unterschied einer (Selbstausfüller-) Online-Befragung gegenüber einer durch Interviewer:innen geführten telefonischen Befragung sofort evident.

beide Richtungen sucht. Solche Ansätze, die Wissenschaftskommunikation dialogisch und partizipativ denken – im angelsächsischen Raum auch als *Science Engagement* oder *Public Engagement* bezeichnet – gibt es in Veranstaltungen der Wissenschaftskommunikation in einigen Beispielen, allerdings nicht in der Fläche. Bisher gibt es auch nur wenig Wirkungsuntersuchungen zu partizipativen Ansätzen. Welche Ziele verfolgen sie? Welche Ergebnisse und welchen Nutzen können sie haben? Wie sollten sie durchgeführt werden, um Wissenschaftler:innen und Bürger:innen gleichermaßen anzusprechen?

Vor diesem Hintergrund entstand bei der Robert Bosch Stiftung die Idee, selbst eine Veranstaltung zu konzipieren mit dem Ziel, einen gleichberechtigen Dialog zwischen Bürger:innen und Wissenschaftler:innen zu initiieren und auf diese Weise gegenseitiges Verständnis und Vertrauen zu fördern. Forschende sollten ermutigt werden „rauszugehen" und insbesondere auch mit Menschen ins Gespräch zu kommen, die sonst kaum Berührungspunkte zu Wissenschaft haben.

Zunächst wurde eine Fokusgruppe ins Leben gerufen, die aus Vertreter:innen beider Gruppen bestand. So testeten z. B. ein Käsethekenverkäufer, eine Hausfrau, ein Materialwissenschaftler und eine Zoologin Fragestellungen und Inhalte der geplanten Dialogveranstaltung und gaben gute Hinweise auf mögliche Fallstricke und Herausforderungen.

2.2 Die Pilotveranstaltung „Mensch Wissenschaft!" im Herbst 2018 in Essen

Auf den Erkenntnissen aus der Fokusgruppe aufbauend lud die Robert Bosch Stiftung im Herbst 2018 eine größere Gruppe von Bürger:innen und Wissenschaftler:innen zu Gesprächen und gleichberechtigtem Austausch in die Zeche Zollverein nach Essen ein. Ziel der Veranstaltung war es, Wissenschaftler:innen dafür zu sensibilisieren, dass der Austausch über ihre Arbeit mit der Gesellschaft notwendig ist und wie das erfolgreich gestaltet werden kann. Gleichzeitig soll Bürger:innen die Möglichkeit geboten werden, Berührungsängste abzubauen und den Zugang zu wissenschaftlichen Erkenntnissen zu erleichtern.

Die Forschenden wurden unabhängig von ihrer Fachexpertise eingeladen. Sie vertraten die Wissenschaft als System und sollten auch selbst zuhören und Fragen stellen: Was bewegt die Menschen? Was wissen sie über Wissenschaft? Welches Bild haben sie von wissenschaftlichen Einrichtungen? Bei der Auswahl wurde auf eine möglichst große Vielzahl von Fächern Wert gelegt, um die Diversität der wissenschaftlichen Forschungsgebiete zumindest im Ansatz zu zeigen.

Umgekehrt sollten die Bürger:innen die Mehrzahl der Teilnehmenden stellen und einen vielfältigen Querschnitt der Bevölkerung widerspiegeln. Dazu wurden zufällig ausgewählte Menschen mit verschiedenen Bildungshintergründen, Nationalitäten, Altersstufen und Milieus angerufen und angeschrieben. Ziel war es, eine gute Mischung der Gruppe mittels verschiedener Kriterien zu erreichen. Alle Teilnehmenden kamen aus Essen oder dem angrenzenden Umland. Zwei Tage verbrachten sie in verschiedenen Konstellationen und zu unterschiedlichen Fragestellungen miteinander, hörten einander zu und teilten sich mit.

2.3 Anforderungen an die Evaluation

Geplant war, aus dem Projekt einen Werkzeugkasten zu entwickeln, der Akteur:innen der Wissenschaftskommunikation zur Verfügung gestellt werden sollte, damit auch an anderen Orten erfolgreiche Veranstaltungen dieser Art umgesetzt werden könnten. Für Bürger:innen sollte das System Wissenschaft in seinen vielen Facetten erlebbar werden, und umgekehrt sollten Wissenschaftler:innen die Möglichkeit erhalten, sich in der Wissenschaftskommunikation zu engagieren.

Um die Projektgenese und das Veranstaltungskonzept auszuwerten und weiterzuentwickeln, war eine begleitende Evaluation eingeplant. Beabsichtigt war jedoch keine Evaluation im klassischen Sinne, die das Dialogexperiment anhand zuvor festgelegter Kriterien überprüfen sollte. Das Institut com.X sollte vielmehr den gesamten Prozess – von der Veranstaltungskonzeption über die Durchführung bis hin zur Kommunikation im Nachgang – mit vielfältigen Methoden kritisch begleiten, die Zielerreichung evaluieren und Empfehlungen zur Anpassung des Formats nach der Pilotveranstaltung ausarbeiten. Begleitend nahm das Projektteam während der Veranstaltung vor allem die Rolle als „kritischer Beobachter“ ein und notierte alles, was an Schwächen und Stärken der Veranstaltung oder Rückmeldungen der Teilnehmenden zu Tage kam. Ziel war es, auf diese Weise im Sinne eines „lernenden Projekts“ ein neuartiges Dialogformat zu entwickeln, das die Robert Bosch Stiftung in einer zweiten, weiterentwickelten Ausgabe 2019 erneut durchführen würde. Nach dieser zweiten Ausgabe sollte das Format Hochschulen und anderen Akteur:innen aus der Wissenschaft als „Blaupause“ für die Ausrichtung ähnlicher Dialogveranstaltungen zur Verfügung gestellt werden (Robert Bosch Stiftung GmbH 2020).

2.4 Die zweite Veranstaltung „Mensch Wissenschaft!“ im Herbst 2019 in Stuttgart

Mit den in Essen gesammelten Erfahrungen hat das Projektteam der Stiftung „Mensch Wissenschaft!“ im November 2019 erneut ausgerichtet – dieses Mal in Stuttgart. Wieder trafen an zwei Tagen viele spannende Persönlichkeiten und Lebensläufe aus Wissenschaft und Gesellschaft in ungewöhnlichen Gesprächsformaten aufeinander. Etwa 25 Wissenschaftler:innen erhielten die Gelegenheit zum Austausch mit rund 50 zufällig und vielfältig ausgewählten Bürger:innen. Ein besonderer Fokus lag auf dem persönlichen Kontakt und der Durchmischung der Gruppen. Formate wie soziometrische Aufstellungen, Kennenlern-Triaden, Kleingruppenarbeiten und themenzentrierte Dialogräume bis hin zu gemeinsamer Prozess-Reflexion und ausreichend Zeit für informelle Gespräche, z. B. beim Essen, boten einen professionellen und bereits bewährten Rahmen für gelungenen Austausch. Wieder hat das Institut com.X die Veranstaltung von Beginn an begleitet und verschiedene Befragungen (vor Ort, online und telefonisch) mit den Teilnehmenden zu Verlauf und Eindrücken durchgeführt.

3 Design der Evaluation

3.1 Formativer Begleitansatz

Obwohl die Planung der Pilotveranstaltung 2018 in Essen zum Zeitpunkt der Beauftragung der externen Evaluation bereits begonnen hatte, erfolgte die Begleitung durch com.X, wie oben beschrieben, aus einer stark formativ geprägten Evaluationsrolle[9] heraus (siehe auch Volk in diesem Band). Anstelle einer rein zielorientierten Wirkungsanalyse galt es vielmehr, Gelingensfaktoren und förderliche Rahmenbedingungen für einen konstruktiven Dialog von Wissenschaftler:innen und Bürger:innen in den Blick zu nehmen.

Entscheidend für diese formative Begleitung war von Anfang an ein enger Austausch mit dem gesamten Projektteam aufseiten der Stiftung. So wurden

[9] Im Gegensatz zur summativen Evaluation, die meist auf Zielerreichung und Wirkungsmessung fokussiert und dadurch oft grundlegende Entscheidungen über die Fortsetzung von Maßnahmen und Programmen vorbereitet, konzentriert sich die meist ihren Gegenstand begleitende formative Evaluation eher auf Prozesse und trägt entsprechend zu einer optimalen Gestaltung bei (siehe auch Volk in diesem Band).

in einer Phase, in der Struktur und Inhalte der Veranstaltung noch nicht fixiert waren, durchaus auch unterschiedliche interne Perspektiven auf das Format sichtbar: Etwa zur Zielgruppenerreichung, (intendierten) Effekten, Akzeptanz und/oder Optimierungsbedarf für Programmelemente und Gesamtveranstaltung. Durch viele persönliche Gespräche unmittelbar vor, während und nach der Veranstaltung wurde der Blick der Evaluation so auch auf bisher unberücksichtigte Aspekte gelenkt, wie bspw. dem Prozess der Ansprache und Gewinnung von Teilnehmenden oder dem möglichen Einfluss des Ortes, seiner Architektur und damit verbundenen Konnotationen auf die Wahrnehmung der Veranstaltung (mehr dazu siehe unten). Mit den Vorbereitungen zur zweiten Ausgabe in Stuttgart ging die Methoden-/Designentwicklung dann „Hand in Hand" mit der Formatkonzeption, zum Beispiel durch eine Teilnahme an einem Partner:in-/Dienstleistende-Workshop der Robert Bosch Stiftung, um die Evaluation frühzeitig zu integrieren in Konzeption, Umsetzung und Nachbereitung der Veranstaltung, insbesondere im Hinblick auf eine Ergebnisbroschüre für potenzielle Interessent:innen bzw. Umsetzende vergleichbarer Dialogveranstaltungen.

Neben dem Anspruch, aus einer neutralen und unabhängigen Perspektive heraus eigene Ergebnisse und Impulse für die Weiterentwicklung zu generieren, hatte die externe Evaluation durch com.X auch die Aufgabe, die Erkenntnisse der zahlreichen selbstevaluativen Schritte zu integrieren, die die Stiftung während und nach der Veranstaltung umsetzte (siehe auch unten). An dieser Stelle soll aber zunächst auf das methodische Design der externen Evaluation eingegangen werden.

3.2 Einzelmethoden und deren Zusammenspiel

Das folgende Schaubild zeigt zunächst die in beiden Jahren eingesetzten Methoden, bevor diese im Weiteren sowohl einzeln als auch im Zusammenspiel beschrieben werden (Abb. 1):

Das *Desk Research* – verbunden mit telefonischen Vorfeld- und Nachbereitungsgesprächen mit dem Projektteam der Stiftung – diente zunächst der Generierung einer detaillierten Wissensbasis zur Konzeption und geplanten Umsetzung der Veranstaltung. Dazu gehörten u. a. auch Hintergründe zur Genese der Formatidee. So wurde Mensch Wissenschaft! in einem ca. einjährigen projektinternen Prozess unter Einbeziehung externer Fachakteur:innen aus Wissenschaftskommunikation und Moderation sowie von Ergebnissen aus Gruppendiskussionen mit Bürger:innen entwickelt. In Verbindung mit dem seitens der Verantwortlichen formulierten Handlungsbedarf (Warum braucht es

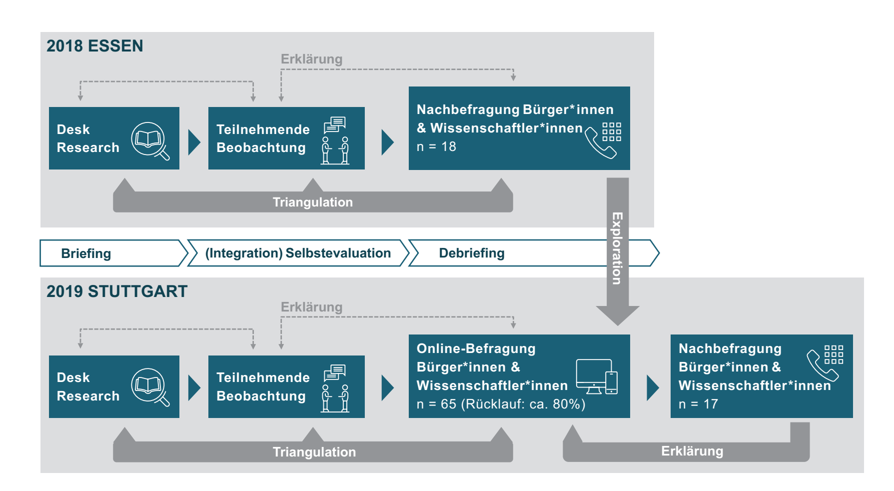


Abb. 1 Methodisches Design der Evaluation. (Quelle: eigene Darstellung, com.X Institut)

dieses Format und wie soll es sich von bestehenden abgrenzen?) und der Verortung innerhalb der Gesamtaktivitäten der Robert Bosch Stiftung im Bereich Wissenschaft und Gesellschaft wurde der Kontext sichtbar, in dem Mensch Wissenschaft! entstanden ist. Diesen galt es aus Evaluationssicht stets sorgfältig zu erfassen, um auch die ursprünglichen Intentionen hinter den formulierten Zielen und Maßnahmen nachvollziehen zu können.

Ein besonderes Augenmerk legte das Desk Research auf die Analyse der Zusammensetzung der Teilnehmenden und den Prozess der Teilnehmendenansprache und -gewinnung. Mensch Wissenschaft! verfolgte den Anspruch, aufseiten der Bürger:innen eine möglichst heterogene, mehr an soziodemografischer Diversität als Repräsentativität orientierte Teilnehmendenschaft zu erreichen. Eine entsprechende Auswertung der Teilnehmendenliste zeigte in einem ersten Schritt, inwieweit dies faktisch gelungen ist.[10] Jedoch konnte erst über die Beobachtung, Interviews und Onlinebefragung geklärt werden, inwieweit mit diesem Ansatz auch wissenschaftsferne oder sogar -kritische Personen erreicht wurden.

[10] Zudem ließ sich für die Folgeveranstaltung in Stuttgart nachvollziehen, inwieweit die (sich bei Online-Befragungen ja selbstselektiv bildende) Stichprobe der quantitativen Nachbefragung der Teilnehmendenschaft entspricht, was wiederum bei der Interpretation der Ergebnisse berücksichtigt wurde.

Die weitere Auseinandersetzung mit dem gesamten Rekrutierungsprozess auf Grundlage von Dokumenten und Gesprächen mit dem Projektteam der Stiftung wie auch dem beauftragten Dienstleister[11] lenkte den Blick der Evaluation in diesem Zusammenhang auch auf die Bedeutung bzw. Limitierung extrinsischer Motivationsanreize (Bürger:innen erhielten für ihre Teilnahme eine Aufwandsentschädigung von 100 € in Essen bzw. 50 € in Stuttgart).

Eine *teilnehmende Beobachtung* (siehe auch Weiß in diesem Band) – orientiert an einer qualitativen Checkliste – erfolgte an allen Veranstaltungstagen durch die Evaluationsverantwortlichen, wodurch auch persönliche Eindrücke von der Mehrheit der Arbeitsgruppen gewonnen werden konnten. Auch wenn dabei (im Sinne der Beobachtung) auf eine aktive Rolle, bspw. durch eine Beteiligung an Diskussionen verzichtet wurde, wurde sowohl in der allgemeinen Begrüßung zu Beginn wie auch auf Nachfrage oder in Vorstellungsrunden die eigene Evaluationsrolle stets transparent gemacht. Letztlich deckt sich hier der eigene Eindruck aber mit den Erfahrungen aus zahlreichen weiteren Veranstaltungsbegleitungen, demnach Evaluationsvertreter:innen – ähnlich zu Kamera-Teams – nach einer Weile nicht mehr als solche wahrgenommen oder direkt dem/der Veranstalter:in zugerechnet werden.

Abgesehen von grundlegenden Eindrücken zum Verlauf von Arbeitsgruppen und Diskussionen oder zur Akzeptanz erarbeiteter Ergebnisse, die sich durch spontane Kurzgespräche mit Teilnehmenden anreichern lassen und im weiteren Forschungsprozess qualitative wie quantitative Ergebnisse ergänzen und deren Interpretation unterstützen, sind teilnehmende Beobachtungen aber vor allem ein unverzichtbarer Bestandteil im Methodenset, um den atmosphärischen Rahmen der Veranstaltung zu erfassen.

Bereits in den ersten Stunden von Mensch Wissenschaft! lieferte die teilnehmende Beobachtung, häufig den Ort und damit auch sich bildende und mischende Personengruppen wechselnd, für die weitere Evaluation wertvolle Einschätzungen dazu, inwieweit der Distanzabbau zwischen Bürger:innen und Wissenschaftler:innen gelang und welchen Anteil daran insbesondere spezielle

[11] Das resultierende Schema für die telefonische Rekrutierung erwies sich als sehr anspruchsvoll, weshalb einzelne Gruppen (z. B. Jüngere, die nicht mehr per Festnetz erreichbar sind) zunächst schwer erreichbar blieben. Deshalb wurden ergänzend schriftliche Einladungen auf Basis eingekaufter Daten des Einwohnermeldeamtes versandt. Der Ansatz weist Parallelen zu Mixed-Mode-Befragungen auf, in denen ebenfalls unterschiedliche Ansprachewege oder Methoden kombiniert werden, um soziodemografische Verzerrungen einzelner Ansätze auszugleichen.

Moderations- und Kennenlerntechniken hatten, die als „Eisbrecher“ zum Einsatz kamen.

Auch die Bedeutung des Veranstaltungsortes für die Wahrnehmung eines Formats erschließt sich unmittelbar stark durch die teilnehmende Beobachtung, ergänzt durch die Einschätzungen der Teilnehmenden in Interviews und Befragungen. Neben eher formalen Rahmenbedingungen wie Raumgrößen und akustischen Verhältnissen können auch die Architektur (wie beim SANAA-Gebäude in Essen mit seiner ungewöhnlichen Rohbetonfassade mit den großen Fensterflächen) oder die sonstige/bisherige Nutzung eines Ortes eine Rolle spielen. So wurde das Stuttgarter Stadtpalais, in dem die Folgeveranstaltung stattfand, auch schon deshalb sichtbar gut von den teilnehmenden Bürger:innen angenommen, da es einem Großteil bereits in seiner früheren Funktion als Zentralbibliothek bekannt war.

Im Nachgang beider Veranstaltungen (Essen/Stuttgart) wurden *qualitative telefonische Leitfadeninterviews* (zu qualitativen Befragungen allgemein siehe auch Metag und Scheu in diesem Band) mit Bürger:innen und Wissenschaftler:innen geführt. Da das grundsätzliche Einverständnis zur Evaluationsteilnahme und Kontaktierung bereits vorab im Zuge des Registrierungsprozesses eingeholt worden war, war es möglich, aus der bereitgestellten Teilnehmendenliste eine breite und diverse Stichprobe zu ziehen, die sich an der soziodemografischen Verteilung des Teilnehmendenfeldes orientierte.

Die leitfadengestützten Interviews boten durch offene Impuls- und Nachfragen den Teilnehmenden viel Raum, ihre Teilnahmeerfahrung, erarbeitete Ergebnisse, gebotene Rahmenbedingungen sowie Programm und Ablauf der Veranstaltung zu reflektieren. Besondere Bedeutung im Hinblick auf die Ziele der Pilotumsetzung hatte dabei die Identifikation von förderlichen wie auch hemmenden Faktoren für einen konstruktiven und wertschätzenden Dialog auf Augenhöhe.

Die Herausforderung in der Analyse bestand vor allem darin, auf Grundlage qualitativ gewonnener Erkenntnisse zwischen individuellen, kontextabhängigen Teilnahmeerfahrungen und allgemeingültigen und damit übertragbaren Erkenntnissen zu differenzieren. Da im Nachgang der Pilotveranstaltung in Essen keine quantitative Befragung erfolgte (wie im Folgejahr in Stuttgart), spielte dafür der Abgleich mit den Ergebnissen der teilnehmenden Beobachtung (alle Interviews wurden von Beobachter:innen geführt) wie auch den Wahrnehmungen des Projektteams (inklusive der Ergebnisse der durch das Projektteam teils in die Veranstaltung eingebetteten selbstevaluativen Elementen) selbst eine große Rolle. Weitere qualitative Interviews mit Expert:innen für Beteiligungsformate, die die Veranstaltung vor dem Hintergrund eigener Erfahrungen begleiteten, unterstützten die Ableitung möglichst verallgemeinerbarer Ergebnisse.

Letztlich trägt der Verzicht auf eine quantitative Erhebung in Essen aber auch dem explorativen Charakter der Pilotumsetzung von Mensch Wissenschaft! Rechnung. Die eher wie offene Gespräche geführten Interviews lenkten den Blick der Evaluation so auf Themen, die zuvor keine große Rolle im Programmkonzept spielten, sich aber als hoch relevant für die Veranstaltungsbewertung erwiesen (wie bspw. den Wunsch von Bürger:innen, mehr über den auch „banalen Alltag" von Wissenschaftler:innen zu erfahren). Diese Hinweise flossen dann nicht nur in die Konzeption der Stuttgarter Veranstaltung selbst, sondern auch – im Sinne einer Exploration – in die dort eingesetzte Online-Befragung ein (z. B. in Form von Aussagebewertungen) und konnten damit im Sinne eines zeitlich mehrere Umsetzungen umfassenden, aufeinander aufbauenden multimethodischen Designs ein Jahr später auch quantitativ validiert werden. Die Interviews mit Teilnehmenden aus Stuttgart fanden dann im Nachgang der Online-Befragung statt und dienten vor allem der qualitativen Vertiefung und Analyse von Hintergründen, Bewertungszusammenhängen und Motivationen.

Die *Online-Befragung* (siehe auch Böhmert und Abacioglu in diesem Band) in Stuttgart verfolgte mehrere Zielsetzungen, sowohl innerhalb der Evaluation als auch darüber hinaus.

In der engen Verzahnung mit der anschließenden qualitativen Nachbefragung lieferte sie zunächst eine standardisierte Rückmeldung zum Format in der Breite. Im Zuge des Veranstaltungsmanagements und des Bestrebens der Robert Bosch Stiftung, auch weiterhin mit den Teilnehmenden in Kontakt zu bleiben und deren Vernetzung untereinander zu fördern, fragte sie aber konkret auch nach Wegen und Wünschen, frisch geknüpfte Kontakte weiter zu vertiefen.

Für die spätere Kommunikation der Stiftung mit Akteur:innen der Wissenschaftskommunikation (in Form der bereits erwähnten Handreichung zum Format) bot die Befragung aber vor allem Ergebnisse, die sich in ihrer Unterscheidung von Bürger:innen- und Wissenschaftler:innen-Perspektive schnell und intuitiv erfassen lassen und in ihrer Quantifizierung und grafischen Darstellung eine Unmittelbarkeit und damit auch Überzeugungskraft besitzen, die rein qualitativ gewonnenen Erkenntnisse oftmals abgeht.[12] Ein gutes Beispiel dafür ist das im Vorwort der Broschüre (Robert Bosch Stiftung GmbH 2020) zitierte

[12] Auch diese eher „kommunikativen" (und nicht nur rein methodischen) Überlegungen können ein Grund sein, bei Evaluationen quantitative und qualitative Methoden zu mischen, vor allem wenn es etwa darum geht, in Richtung von Entscheider:innen oder Fördermittelgebern zu argumentieren.

Befragungsergebnis, demnach 100 % der (befragten) Bürger:innen und 95 % der Wissenschaftler:innen an ähnlichen Veranstaltungen erneut teilnehmen würden.

3.3 Verschiedene Typen von Mixed-Method-Designs

Die externe Evaluation von Mensch Wissenschaft! kombinierte nicht nur verschiedene empirische Methoden miteinander: Teilnehmende Beobachtung, quantitative Befragung und qualitative Interviews boten im Sinne einer *Triangulation* verschiedene Zugänge zu letztlich stets individuellen Teilnahmeerfahrungen und ermöglichten damit erst die Ableitung verallgemeinerbarer Erkenntnisse für die Weiterentwicklung und Kommunikation zum Format. Der insbesondere bei der Pilotumsetzung in Essen gegebene experimentelle Veranstaltungscharakter bedingte dabei zunächst einen offenen, qualitativen Evaluationsansatz, dessen stark subjektiv geprägte Erkenntnisse dann im Sinne eines *„exploratory design"* über die quantitative Onlinebefragung in Stuttgart validiert und generalisiert wurden. Die darauf folgenden qualitativen Interviews dienten hingegen vor allem der Vertiefung und Klärung *(„explanatory design").*

Zudem bot Mensch Wissenschaft! mit unmittelbar in die Veranstaltung und das Veranstaltungsmanagement integrierten Evaluationselementen aber auch ein gutes Beispiel für eine handhabbare *Selbstevaluation* (Robert Bosch Stiftung GmbH 2020). So dienten Abstimmungsrunden der Projektakteur:innen zwischen Veranstaltungsblöcken nicht nur der weiteren Veranstaltungskoordination, sondern auch der Erfassung systematischer Eindrücke aus unterschiedlichen Akteur:innen-Perspektiven: etwa zur Zielgruppenerreichung, (intendierten) Effekten, Akzeptanz und/oder Optimierungsbedarf für Programmelemente und Gesamtveranstaltung. Darüber hinaus wurden Möglichkeiten für Teilnehmendenfeedback direkt in den Veranstaltungsablauf integriert, etwa über Kartenabfragen oder ein Reflexions-Plenum zum Abschluss.

Die externe Evaluation integrierte diese Erkenntnisse und unterstützte das Projektteam bei der Reflexion und Implikation der geleisteten Arbeit durch eigene Beobachtungen, Gespräche mit Teilnehmenden, der Moderation, Vertretung der Hochschul-Partner:innen oder die Teilnahme an internen Debriefings während der Veranstaltung.

4 Erkenntnisse und Nutzen der Evaluation[13]

4.1 Essener Veranstaltung bietet erkenntnisreichen „Probelauf“

Im Einzelnen wurden bereits während und nach der Veranstaltung in Essen aufseiten der Robert Bosch Stiftung einige Erkenntnisse offenbar, die sich als relevant für einen gelungenen Dialog erwiesen:

- Das ungleiche Verhältnis von 2/3 Bürger:innen zu 1/3 Wissenschaftler:innen hatte zu einer ausgewogenen Diskussionskultur beigetragen. Die gleiche Wirkung hatten Namensschilder ohne Titel, die eine direkte Zuordnung einer Person zu einer der beiden Gruppen nicht ermöglichte. Man begegnete sich auf Augenhöhe und erfuhr erst im Gespräch miteinander, ob der eine Bürger oder die andere Wissenschaftlerin war.
- Den beteiligten Wissenschaftler:innen war teilweise ihre Rolle unklar. Sind sie normalerweise gewohnt, auf Fachkonferenzen Vorträge vor anderen Fachleuten zu halten, waren sie in der Rolle als Laie (gegenüber anderen Disziplinen) oder als Zuhörende verunsichert. Auch Fragen nach der Arbeitsweise von Wissenschaft und dem persönlichen Arbeitsalltag hatten den gleichen Effekt. Auf ihre Rollen als Expert:innen, Laien und Zuhörende sowie auf die Beantwortung von Fragen, die über die reine Fachexpertise hinausgingen und vor allem auch auf die Nutzung einfacher und verständlicher Sprache, hätten sie besser vorbereitet werden müssen.
- Eine konsequente Durchmischung der Gruppen erwies sich als hilfreich für gelungenen Dialog. Insgesamt sollte das Programm einfach und ohne zu viele Vorgaben gehalten werden. Tatsächlicher Dialog entwickelte sich am besten in Kleingruppen oder in Ess- und Kaffeepausen.

Inhaltlich kristallisierten sich neben konkreten gesellschaftlichen Fragen wie bspw. Klimawandel und Ernährung folgende Themen als interessant für die Teilnehmenden heraus: Werteverständnis von Wissenschaft, Kontrolle (durch den Staat), Transparenz, Finanzierung (durch die Industrie), „Glauben“ an Wissenschaft, Fortschritt, Unabhängigkeit, gutes Leben (trotz/dank Wissenschaft),

---

[13] Weitere Details zu den Erkenntnissen und Ergebnissen siehe Robert Bosch Stiftung GmbH (2020).

Gemeinwohlorientierung von Forschung und „Lebensnähe“. Es wurde deutlich, wie wichtig es ist, Wissenschaft verständlich zu erklären. Auch Formate, die wissenschaftliche Grundbildung – sogenannte *Scientific Literacy* – vermittelten, stießen auf großes Interesse.

Das Interesse der Teilnehmenden an mehr Austausch und Engagement auch über die Veranstaltung hinaus sollte von Anfang an mitgedacht werden. Die Wirkung einer einmaligen Veranstaltung geht schnell verloren. Gewecktes Interesse sollte mit Vertiefungsmöglichkeiten oder Nachfolgeveranstaltungen erhalten und gestillt werden.

4.2 Angepasstes Veranstaltungskonzept in Stuttgart

Im Konzept der zweiten Veranstaltung in Stuttgart wurden die oben genannten Punkte sowie weitere Ergebnisse aus der externen Evaluation berücksichtigt und entsprechend eingearbeitet:

- Der didaktische Aufbau der Veranstaltung wurde angepasst. Statt mit übergreifenden Fragen an die Wissenschaft den Dialog zu starten, wurden in Stuttgart zunächst konkrete Forschungsthemen vorgestellt, bevor die Moderation am zweiten Tag in sogenannten „Metaworkshops“ eine Brücke zu abstrakteren Themen (Unabhängigkeit der Wissenschaft, Vertrauen in die Wissenschaft, etc.) schlug.
- Der Wunsch der Essener Teilnehmenden nach mehr Zeit für Diskussionen und ein Resümee im Plenum, führte zu einer deutlichen Entzerrung des Programms der Stuttgarter Veranstaltung. Es gab zwar weiterhin Wechsel zwischen Arbeits- und Diskussionsphasen, zugleich aber genügend Gelegenheiten für informellen Austausch beim gemeinsamen Essen oder in Kaffeepausen. Die Evaluation bestätigte, dass alle Beteiligten Berührungsängste mit jeweils der anderen Gruppe abbauen konnten.
- Die Anpassung der Aufwandsentschädigung (100 € in Essen, 50 € in Stuttgart) trug dazu bei, dass in Stuttgart zwar bildungsfernere Milieus weniger gut erreicht wurden, es dafür aber auch weniger passive bis offen desinteressierte Teilnehmende gab, was sich wiederum förderlich auf die Gesprächskultur auswirkte.
- Zudem wurden den Teilnehmenden in Stuttgart grundsätzlich etwas mehr Anleitung und Impulse geboten. Zwar hatte sich die Selbstorganisation der Diskussionsgruppen in Essen als förderlich für einen Dialog auf Augenhöhe erwiesen (Rollen für Moderation, Präsentation, Dokumentation etc. wurden

zu Beginn selbst verteilt). Trotzdem schuf die Hinzunahme professioneller Moderator:innen in Stuttgart letztlich eine ergebnisorientiertere Struktur, welche die Balance im Dialog wahren und auch hitzige Diskussionen ausgewogen zusammenzufassen konnte.
- Ebenfalls modifiziert wurde, dass sowohl Wissenschaftler:innen wie auch Bürger:innen meist als Laien über (fach-)fremde Themen diskutierten. So wurde den Sessions in Stuttgart ein fachlicher Impulsvortrag vorangestellt, um eine gemeinsame Wissensbasis für einen differenzierteren Dialog zu schaffen. Zudem wurden die Teilnehmenden vorab über die zur Diskussion stehenden Forschungsthemen informiert und konnten sich nach Interesse ihren „Platz" buchen.

Des Weiteren wurden kurze interne ad-hoc-Beratungen in den Verlauf der Veranstaltung eingebaut, in denen das Projektteam der Robert Bosch Stiftung, com.X und die Moderation Rückmeldungen zum Verlauf verschiedener Programmmodule besprachen und kurzfristige Ablaufänderungen direkt umsetzten.

Die begleitende Evaluation hat in hohem Maße zur Verbesserung des Veranstaltungskonzepts beigetragen. Die Ergebnisse flossen in die im Nachgang publizierte Handreichung „Wie Hochschulen Bürger und Wissenschaftler ins Gespräch bringen – Mensch Wissenschaft! Miteinander reden. Voneinander lernen" ein. Die Handreichung stellt den Lernprozess und die Evaluationsergebnisse auf anschauliche Art zur Verfügung. Zielgruppe sind Hochschulen und andere wissenschaftliche Einrichtungen, die mit der Broschüre Möglichkeiten aufgezeigt bekommen, wie sie Aktivitäten im Rahmen ihrer „Third Mission" initiieren können, indem sie z. B. in ihrer Region selbst ähnliche Dialogveranstaltungen ausrichten. Die Broschüre enthält Tipps und Anleitungen zur Selbstevaluation und ist als eine Art „Rezept zum Nachkochen mit Geling-Garantie" konzipiert. Ebenso gibt die Handreichung Beschäftigten der Wissenschaftskommunikation und Institutionsleitungen klare Belege und Entscheidungshilfen an die Hand, ob und wo es sich lohnt, in den Dialog mit der Gesellschaft zu investieren.

In Summe wurden mit „Mensch Wissenschaft!" viele neue persönliche Beziehungen geschaffen, die modellhaft für eine bessere Verwurzelung von Wissenschaft in der Gesellschaft sorgten.

Literatur

Beilharz F, Kattau N, Kratz K, Kopp O, Probst A (2020) Der Online Marketing Manager: Handbuch für die Praxis. O’Reilly, Heidelberg

Burke Johnson R, Onwuegbuzie AJ, Turner LA (2007) Toward a definition of mixed methods research. J Mixed Methods Res 1(2):112–133. https://doi.org/10.1177/1558689806298224

Creswell JW, Plano-Clark V (2007) Designing and conducting mixed methods research. Sage, Thousand Oaks, S 59–67

Gabriel M, Quast T (2006) Gesamtbericht zur Evaluation des Einsteinjahres 2005. Bochum, Hannover. Im Auftrag des BMBF. https://edocs.tib.eu/files/e01fb07/528483862.pdf. Zugegriffen: 21. Juni 2022

Gabriel M, Warthun N (2017) Begleitforschung Wissenschaftsjahr 2015 – Zukunftsstadt. Gesamtbericht. Bochum. Im Auftrag des BMBF. https://www.bmbf.de/bmbf/shareddocs/downloads/files/abschlussbericht_begleitforsch-aftsjahr-2015_finale-fassung-2.pdf;jsessionid=82911B1BED975BD01A84F7729C420D92.live472?__blob=publicationFile&v=1. Zugegriffen: 21. Juni 2022

Kelle U (2015) Mixed methods. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer, Wiesbaden

Krug G, Kriwy P, Carstensen J (2014) Mixed-Mode Designs bei Erhebungen mit sensitiven Fragen: Einfluss auf das Teilnahme- und Antwortverhalten. LASER discussion papers No. 84, Erlangen

Leech NL, Onwuegbuzie AJ (2009) A typology of mixed methods research designs. Qual Quant 43:265–275. https://doi.org/10.1007/s11135-007-9105-3

Plano-Clark V, Huddleston-Casas C, Churchill S, Green N, Garrett A (2008) Mixed methods approaches in family science research. J FAM ISS 29(11). https://doi.org/10.1177/0192513X08318251

Robert Bosch Stiftung GmbH (2020) Wie Hochschulen Bürger und Wissenschaftler ins Gespräch bringen – Mensch Wissenschaft! Miteinander reden. Voneinander lernen. Stuttgart. https://www.bosch-stiftung.de/sites/default/files/publications/pdf/2020-07/Broschuere_Mensch_Wissenschaft_Dialogformat_ES.pdf

Tashakkori A, Teddlie C (1998) Mixed methodology: combining qualitative and quantitative approaches. Sage, Thousand Oaks

Tashakkori A, Teddlie C (2003) Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In: Tashakkori A, Teddlie, C (Hrsg) Handbook of mixed methods in social & behavioral research. Sage, Thousand Oaks, S 3–50

Markus Gabriel ist Partner im com.X Institut für Kommunikations-Analyse und Evaluation. Er ist verantwortlich für Studien und Evaluationen mit breit angelegtem quantitativ-qualitativem Methodenmix. Zu seinen forscherischen Schwerpunkten gehören neben Gesundheits- und Bildungsthemen insbesondere partizipative Prozesse in der Öffentlichkeitsbeteiligung, Regionalentwicklung und Wissenschaftskommunikation.

Isabella Kessel ist Projektleiterin bei Translake. Sie berät Kommunen hinsichtlich Bürgerbeteiligungsverfahren, Strategieentwicklung und Fördermittelakquise und unterstützt kommunale Prozesse in Themen wie z. B. Klimawandel, Mobilität, Digitalisierung, Naturschutz und viele andere. Als Senior Projektmanagerin der Robert Bosch Stiftung im Themenbereich „Wissenschaft in der Gesellschaft" lag der Schwerpunkt ihrer Arbeit in der Förderung von Wissenschaftskommunikation und Wissenschaftsjournalismus. Zu den Zielen gehörten die Institutionalisierung von Public Engagement im Wissenschaftssystem, die Partizipation von Bürger:innen in und an Wissenschaft und der Wissenstransfer in die Gesellschaft.

Thomas Quast ist seit 1999 Geschäftsführer und Gründungsgesellschafter des com.X Instituts für Kommunikations-Analyse und Evaluation. Zu seinen Begleitforschungs- und Evaluationsschwerpunkten gehören Bürger:innenbeteiligung, Wissenschaftskommunikation und andere Wissenstransfer- und Dialogangebote, die den partizipativen Austausch zwischen gesellschaftlichen Subsystemen und Gesellschaft bzw. Bürger:innen ermöglichen und fördern.

Eva Roth ist Projektmanagerin für das Thema Wissenschaft in der Gesellschaft in der Robert Bosch Stiftung. Sie verantwortet Projekte, die sich für neue Entwicklungen und einen offenen Dialog zwischen Wissenschaft, Politik und Gesellschaft einsetzen. Besonders interessieren sie Formate, die die Öffentlichkeit interaktiv in die Forschung einbeziehen sowie Angebote für Menschen, die selten mit Wissenschaft in Kontakt kommen.

